Vers l'intégration de post-éditions d'utilisateurs pour améliorer les systèmes de traduction automatiques probabilistes

Marion Potet

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministérial : 7 août 2006

Présentée par

Marion Potet

Thèse dirigée par **Laurent Besacier**
et codirigée par **Hervé Blanchon**

préparée au sein **du Laboratoire d'Informatique de Grenoble (LIG)**
et de **l'école doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII)**

Vers l'intégration de post-éditions d'utilisateurs pour améliorer les systèmes de traduction automatiques probabilistes

Thèse soutenue publiquement le **9 avril 2013**,
devant le jury composé de :

M. Eric Gaussier
Professeur des Universités à l'Université de Grenoble 1-LIG, Président
Mme. Violaine Prince
Professeur des Universités à l'Université de Montpellier 2-LIRMM, Rapporteur
M. Kamel Smaïli
Professeur des Universités à l'Université de Nancy 2-LORIA, Rapporteur
Mme Lucia Specia
Senior Lecturer à *Sheffield University*, Examinatrice
M. Laurent Besacier
Professeur des Universités à l'Université de Grenoble 1-LIG, Directeur de thèse
M. Hervé Blanchon
Maître de Conférence à Université de Grenoble 2-LIG, Co-Directeur de thèse

Résumé

Les technologies de traduction automatique existantes sont à présent vues comme une approche prometteuse pour aider à produire des traductions de façon efficace et à coût réduit. Cependant, l'état de l'art actuel ne permet pas encore une automatisation complète du processus et la coopération homme/machine reste indispensable pour produire des résultats de qualité. Une pratique usuelle consiste à post-éditer les résultats fournis par le système, c'est-à-dire effectuer une vérification manuelle et, si nécessaire, une correction des sorties erronées du système. Ce travail de post-édition effectué par les utilisateurs sur les résultats de traduction automatique constitue une source de données précieuses pour l'analyse et l'adaptation des systèmes. La problématique abordée dans nos travaux s'intéresse à développer une approche capable de tirer avantage de ces retro-actions (ou post-éditions) d'utilisateurs pour améliorer, en retour, les systèmes de traduction automatique. Les expérimentations menées visent à exploiter un corpus d'environ 10 000 hypothèses de traduction d'un système probabiliste de référence, post-éditées par des volontaires, par le biais d'une plateforme en ligne. Les résultats des premières expériences intégrant les post-éditions, dans le modèle de traduction d'une part, et par post-édition automatique statistique d'autre part, nous ont permis d'évaluer la complexité de la tâche. Une étude plus approfondie des systèmes de post-éditions statistique nous a permis d'évaluer l'utilisabilité de tels systèmes ainsi que les apports et limites de l'approche. Nous montrons aussi que les post-éditions collectées peuvent être utilisées avec succès pour estimer la confiance à accorder à un résultat de traduction automatique. Les résultats de nos travaux montrent la difficulté mais aussi le potentiel de l'utilisation de post-éditions d'hypothèses de traduction automatiques comme source d'information pour améliorer la qualité des systèmes probabilistes actuels.

Mots-clés : traduction automatique probabiliste, apprentissage, évaluation, post-édition humaine, collecte de corpus, post-édition statistique, estimation de confiance.

Abstract

Nowadays, machine translation technologies are seen as a promising approach to help produce low cost translations. However, the current state of the art does not allow the full automation of the process and human intervention remains essential to produce high quality results. To ensure translation quality, system's results are commonly post-edited : the outputs are manually checked and, if necessary, corrected by the user.

This user's post-editing work can be a valuable source of data for systems analysis and improvement.

Our work focuses on developing an approach able to take advantage of these users' feedbacks to improve and update a statistical machine translation (SMT) system.

The conducted experiments aim to exploit a corpus of about 10,000 SMT translation hypotheses post-edited by volunteers through a crowdsourcing platform. The first experiments integrated post-editions into the translation model on the one hand, and on the system outputs by automatic post-editing on another hand, and allowed us to evaluate the complexity of the task. Our further detailed study of automatic statistical post-editions systems evaluate the usability, the benefits and limitations of the approach. We also show that the collected post-editions can be successfully used to estimate the confidence of a given result of automatic translation.

The obtained results show that the use of automatic translation hypothese post-editions as a source of information is a difficult but promising way to improve the quality of current probabilistic systems.

Keywords : statistical machine translation, learning, quality evaluation, human post-editing, corpus collection, statitical post-editing, confidence estimation.

Remerciements

Puisque l'occasion m'en est offerte, je tiens à exprimer publiquement ma gratitude à tous ceux qui ont contribué à la réalisation de ce travail.

Je tiens tout d'abord à remercier sincèrement le directeur de cette thèse, M. Laurent Besacier professeur à l'Université Joseph Fourier, pour m'avoir guidé et conseillé tout au long de ce travail. Je lui suis reconnaissante de la patience et l'humanité qu'il a manifestées à mon égard durant cette thèse. Qu'il trouve dans ces quelques lignes l'expression de toute l'affection et le respect que je lui porte.

Mes chaleureux remerciements vont également à M. Hervé Blanchon, maître de conférence à l'Université Pierre Mendès France. Je lui témoigne ma gratitude pour avoir co-dirigé cette thèse. L'intérêt qu'il a porté à mes travaux, son efficacité ainsi que sa disponibilité m'ont été d'une aide précieuse.

Le suivi constant et l'encadrement de qualité qu'ils m'ont tous deux apportés m'ont offert des conditions plus que favorables pour mener à bien cette étude. Je leur suis très reconnaissante d'avoir dirigé ce projet de recherche et de m'avoir donné une formation de première qualité.

Je remercie enfin chaleureusement ceux qui se reconnaitront et sans qui cette thèse ne serait pas ce qu'elle est...

« La machine conduit l'homme à se spécialiser dans l'humain. »

Jean Fourastié, "Le grand espoir du XXe siècle" (1972)

Table des matières

Table des figures

Liste des tableaux

Introduction

Les tendances actuelles montrent les besoins de plus en plus important en traduction automatique : la croissance rapide d'Internet, l'internationalisation politique et la mondialisation économique augmentent l'abondance des ressources multilingues et le besoin en traduction de textes. Il est alors plus que jamais nécessaire de briser les barrières linguistiques afin de permettre des flux d'échanges multilingues rapides et efficaces.

La qualité des systèmes de traduction automatique a été significativement améliorée durant ces dernières années mais celle-ci n'a pas atteint l'idéal d'une traduction entièrement automatique et de haute qualité espérée lors de l'apparition de la technologie, soixante ans auparavant. Même si les systèmes automatiques permettent d'automatiser (du moins partiellement) la tâche de traduction et de gagner du temps, la qualité des résultats reste souvent bien en deça de celle produite par les traducteurs humains professionnels.

Malgré cela, l'émergence et la diffusion des outils de traduction automatique auprès du grand public (entre autres sur le World Wide Web) fait émerger de nouvelles problématiques et de nouvelles perspectives.

Bien que l'enjeu majeur reste l'amélioration des systèmes de traductions actuels, les acteurs du domaine se concentrent maintenant sur des objectifs plus modestes et cherchent, par exemple, à prédire la difficulté d'une traduction, évaluer au mieux un résultat de traduction ou encore à détecter et analyser les erreurs commises par le système. Ces objectifs sont fortement liés à la problématique de l'amélioration de la traduction en faisant en sorte que les systèmes de traduction soient capables d'anticiper les difficultés, d'en rendre compte de manière intelligible et surtout d'auto-évaluer avec précision la qualité de leurs traductions, voire de solliciter l'aide de l'utilisateur pour améliorer la traduction proposée.

Ces dernières années, l'essor du Web collaboratif a engendré de nouveaux usages grâce, entre autres, à des outils collaboratifs où la participation des internautes/utilisateurs contribue à enrichir les systèmes. C'est ainsi que, dans le domaine du traitement automatique du langage naturel, est apparu récemment l'idée de pallier les défauts actuels des systèmes de traduction en faisant appel aux utilisateurs eux-mêmes.

L'essentiel du travail présenté dans ce manuscrit porte sur la problématique de l'utilisation de rétroactions ou retours d'utilisateurs (*feedback* en anglais) afin d'améliorer un système de traduction automatique.

Le scénario envisagé, décrit dans la figure 1, est de proposer un outil de traduction automatique qui collecte des corrections de résultats faites par les utilisateurs de l'outil. Ainsi collectées, ces rétroactions seraient ensuite utilisées comme source d'amélioration pour permettre au système de traduction automatique de s'adapter de façon interactive et itérative, au fur et à mesure des interactions avec l'utilisateur.

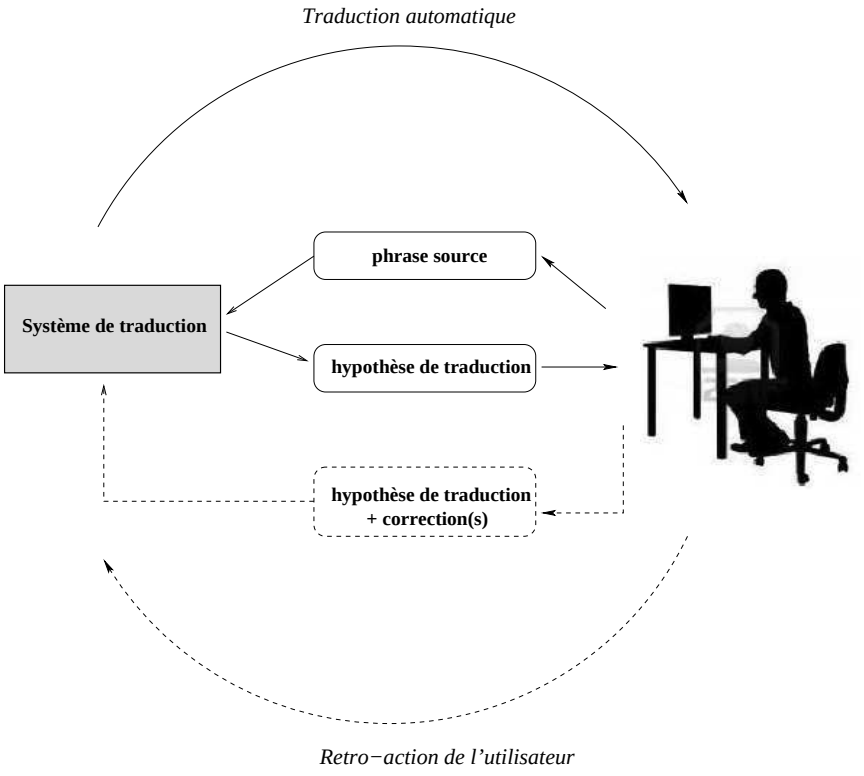


FIGURE 1 – Scénario applicatif de traduction automatisée collaborative

Une application possible de ce scénario est, par exemple, un système de traduction automatique probabiliste (V_i) en ligne sur le Web, qui propose aux utilisateurs d'entrer un texte à traduire et qui offre la possibilité de corriger l'hypothèse de traduction fournie par le système. Cette correction de traduction serait enregistrée et utilisée pour construire une nouvelle version améliorée du système (V_{i+1}).

Au délà de cette application, où la contribution volontaire des internautes est nécessaire, ce scénario peut également être appliqué au contexte de traduction professionnelle. En effet, certains logiciels avec lesquels travaillent les traducteurs professionnels intègrent des outils de traduction automatique utilisés pour obtenir une ébauche de traduction ensuite éditée puis corrigée par le traducteur. Les hypothèses de traductions ainsi corrigées pourraient être utilisées pour améliorer le système de traduction utilisé pour les générer.

Dans le contexte de ce scénario applicatif, le travail de recherche présenté dans ce manuscrit a pour objectif de recueillir des données représentatives de corrections de résultats de systèmes de traduction automatique et de proposer des méthodes visant à enrichir et améliorer le système les ayant produites.

Dans un premier temps, ce manuscrit introduira les éléments fondamentaux nécessaires à la compréhension des travaux présentés : la traduction automatique probabiliste, les modèles de référence, les techniques liées à son évaluation et les approches où l'expertise humaine supplée les systèmes (chapitre I).

Pour les besoins de nos expérimentations, nous avons créé un système de traduction automatique probabiliste à l'état de l'art. Ce système sera considéré comme le système de référence pour la suite. La présentation des corpus utilisés, l'apprentissage du modèle de référence et son l'évaluation feront l'objet d'une deuxième partie (chapitre II).

Le scénario applicatif envisagé nous a conduit à réaliser un premier ensemble d'expérimentations préliminaires destinées à évaluer le potentiel et la faisabilité de l'approche envisagée. Les expériences présentées dans cette troisième partie visent à utiliser les corrections manuelles d'un petit corpus de 175 hypothèses de traductions pour modifier le système à trois niveaux différents du processus de traduction (chapitre III).

Les résultats de ces travaux préliminaires ont fait émerger la nécessite de disposer d'un large corpus d'hypothèses de traductions étiquetées avec leurs corrections. Pour ce faire, nous avons fait appel à des internautes volontaires rémunérés pour corriger un ensemble de 10 000 hypothèses de traduction produites par notre système (chapitre IV).

Nous proposons, par la suite, des solutions visant à exploiter ce corpus de 10 000 post-éditions collectées afin d'améliorer la qualité des résultats du système de traduction automatique de référence (chapitre V). Les résultats obtenus lors de ces travaux nous amèneront ensuite à l'étude de systèmes de post-édition automatique probabilistes de leurs usages jusqu'à leurs limites (chapitre VI).

Dans une dernière partie, nous serons tout naturellement amenés à évaluer l'utilisabilité du corpus collecté à des fins de développement de mesures de confiance pour la traduction automatique (chapitre VII).

Nous conclurons ce manuscrit par une discussion des travaux et résultats présentés et nous présenterons les perspectives de recherche qu'ils permettent d'envisager.

Chapitre I

Etude bibliographique

1 Introduction à la traduction automatique probabiliste

Ce chapitre présente un état de l'art de l'évolution de la traduction automatique probabiliste de ses débuts jusqu'à l'utilisation des modèles log-linéaires. Après un bref historique de la traduction automatique, je présente les deux principales approches : les méthodes expertes et les méthodes empiriques. Les concepts de référence pour la traduction automatique statistique font l'objet d'une troisième section. J'y détaille les fondements théoriques, les notions de corpus et d'alignement, et j'y introduis les modèles à base de mots, ceux à base de segments, puis les différentes mesures d'évaluation de la traduction. La quatrième partie est consacrée aux modèles log-linéaires pour la traduction automatique probabiliste et présente une vue des avancées scientifiques et techniques les plus récentes du domaine.

Dans un premier temps, il s'agit de définir ce que l'on appelle traduction automatique : la traduction automatique désigne, au sens strict, le fait de traduire entièrement un texte grâce à un ou plusieurs programmes informatiques, sans qu'un traducteur humain n'ait à intervenir. Le but est de traduire un texte depuis une langue source, vers une langue cible. Ce projet est relativement ambitieux car le langage naturel est un objet particulièrement complexe.

1.1 Historique de la traduction automatique

L'idée de faire traduire un texte automatiquement fut émise pour la première fois le 4 mars 1947, dans un courrier du scientifique Warren Weaver [1] à son collègue, Norbert Wiener [2](un extrait du courrier est présenté dans la figure I.1). Cette lettre, qui pourrait être à l'origine des efforts de recherche fait en traduction automatique, suggérait

1. Warren Weaver [1894-1978] est un mathématicien Américain. En 1947, il est directeur de la division « Sciences Naturelles » à la fondation Rockefeller.

2. Norbert Wiener [1894-1964] est un mathématicien américain, théoricien et chercheur en mathématiques appliquées.

d'appliquer les techniques bien connues de cryptographie (l'utilisation de systèmes à crypter et dé-crypter datant de l'Antiquité) à des fins de traduction et qu'une machine pourrait être utilisé à ce sujet.

> *[...] One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley*[a] *has recently told me that they are appalled by the magnitude and the importance of the translation job. Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (were the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while. [...]*
>
> *a.* Julian Huxley est, en 1947, premier directeur de l'UNESCO.

FIGURE I.1 – Extrait d'une lettre de Warren Weaver à Norbert Wiener, 4 mars 1947.

C'est lors de la Seconde Guerre mondiale que les premiers essais d'élaboration de machines à traduire automatiquement des messages voient le jour. L'armée américaine tente alors de mettre au point des systèmes susceptibles de déchiffrer les messages codés de l'armée japonaise. La sortie des premiers ordinateurs, en 1949 à l'Université de Washington, donna de nouvelles perspectives à l'idée de traduction automatique.

Les tentatives de traduction automatique se sont poursuivies et développées, principalement aux États-Unis et en Union soviétique, durant la guerre froide. La traduction automatique se contente alors d'une traduction mot à mot, complétée par une base de données d'expressions courantes. Cette première méthode sera ensuite enrichie par des modules métiers spécifiques (informatique, santé, militaire). C'est d'ailleurs à cette époque qu'émergent en plein contexte de guerre froide, les sociétés Systran et Prompt, aujourd'hui encore aux premières loges. La première est américaine, et travaille pour la CIA afin d'assurer la traduction automatique de textes russes vers l'anglais. La seconde, de nationalité russe, effectue le travail inverse pour le compte de son pays.

À des recherches réalistes qui visaient à mettre au point des dictionnaires automatiques bilingues qui simplifiaient le travail de traducteurs humains, a succédé une période d'enthousiasme dans les années 1950-1960, où, grâce notamment au développement des techniques informatiques, plusieurs projets concurrents de traduction ont vu le jour aux États-Unis et en Europe. En 1955 est édité le premier ouvrage dédié à la traduction automatique [Lock et Booth 1955] et la première conférence internationale sur le sujet a lieu en 1956, au MIT [3]. Face aux progrès rapides de la traduction automatique, Y. Bar-Hillel (chercheur du domaine à temps plein depuis 1951 au MIT) publia en 1960 un premier article contenant une mise en garde contre les attentes trop

3. Massachusetts Institute of Technology.

élevées vis à vis de cette nouvelle technologie en mettant en évidence les limitations théoriques de la traduction automatique [Bar-Hillel 1960] : selon lui, certains exemples de traduction sont hors de portée des machines car ils exigent une connaissance globale du monde (voir figure I.2). La faisabilité de la traduction « entièrement » automatique est alors mise en doute.

The box is in the pen.

Peut être traduit en français par :
- « La boîte est dans le parc » ou ;
- « La boîte est dans le stylo ».

The pen is in the box.

Peut être traduit en français par :
- « Le stylo est dans la boîte » ou ;
- « Le parc est dans la boîte ».

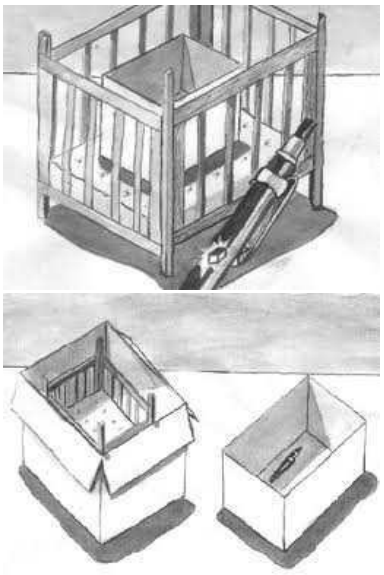


FIGURE I.2 – Exemples de problème de traduction d'une phrase contenant des mots polysémiques, extraits de [Bar-Hillel 1960].

C'est en 1966 que la recherche en traduction automatique fut considérablement ralentie suite à la parution d'un rapport du comité ALPAC (pour *Automatic Language Processing Advisory Committee* en anglais) demandé par le gouvernement américain. Ce rapport [4] [Pierce et al. 1966] proposait une conclusion très sceptique sur le résultat des recherches faites jusqu'à présent en traduction automatique et suggérait de centrer les recherches sur d'autres aspects de la linguistique computationnelle. Suite à ce rapport, le gouvernement Américain réduit considérablement les financements dédiés aux projets de traduction automatique.

Relancées au début des années 80, les recherches sur la traduction automatique utilisent alors les idées de l'intelligence artificielle et la puissance de traitement accrue des nouvelles générations d'ordinateurs. Toutefois, très rapidement, on a pu constater que les objectifs visés étaient irréalisables, et les progrès insignifiants. Plusieurs projets ont ainsi été abandonnés, d'autres sévèrement limités, soit dans leur champ d'application (plusieurs programmes se contentent de chercher à traduire des textes très simples, appartenant à un domaine précis), soit dans leur nature (plusieurs programmes sont complétés par une post-édition assurée par un correcteur-traducteur humain). La recherche est à nouveau mise en veille au profit d'applications moins ambitieuses telles que la Traduction Assistée par Ordinateur.

4. http ://www.nap.edu/openbook.php ?record_id=9547

Il faudra attendre la fin des années 90 et l'avènement de la société de l'information pour voir apparaître de véritables débouchés aux logiciels de traduction automatique : traduction à la volée de pages Web, traduction de dépêches d'actualité, etc. Lorsque ces logiciels quittent les environnements « grands systèmes » pour s'installer sur le poste client, la traduction automatique s'enrichie d'une dizaine de paires de langues supplémentaires et d'une analyse statistique.

Depuis quelques années, les solutions de traduction automatique connaissent un essor considérable grâce au développement et à l'expansion des réseaux de communication (Internet, technologies Web, téléphones intelligents, etc.). La traduction automatique par ordinateur, longtemps réservée à l'usage des services publics (défense, gouvernement, administration), a alors gagné les entreprises et les particuliers sous l'impulsion des progrès en informatique et de la mondialisation culturelle et économique.

Sur le World Wide Web en particulier, la prolifération des données textuelles et des outils gratuits font naître de nouveaux besoins. La traduction automatique en ligne est relativement récente puisque le premier outil de traduction automatique mis gratuitement en service sur le web a été un système de Systran en 1997 (Babel Fish sur AltaVista). Aujourd'hui, divers systèmes en ligne permettent de traduire automatiquement des pages Web et des textes plus ou moins brefs. De tels outils présentent plusieurs autres avantages, pour les utilisateurs : ils sont gratuits (pour la plupart), ils sont « instantanés », acceptent des entrées de différentes natures (mots, textes, mail, page web, etc) et proposent la traduction de plusieurs langues.

A l'heure actuelle, cette technologie est appréciée du grand public car elle permet de pouvoir comprendre le thème d'un texte dans une langue totalement inconnue et les principaux faits ou éléments d'information qu'elle contient.

1.2 Fonctions et usages de la traduction automatique

On distingue traditionnellement deux usages de la traduction automatique : l'assimilation d'information et la dissémination, ou diffusion, d'information. La traduction automatique est mise en oeuvre dans un but d'assimilation quand une personne utilise le système pour comprendre l'information contenue dans un document qui est écrit dans une langue qu'elle ne connait pas ou peu. C'est, par exemple, l'usage principal que les internautes font des traducteurs disponibles en ligne. Un système de traduction automatique est utilisé dans un but de dissémination quand son résultat est destiné, après vérification par un spécialiste de la langue, à être diffusé à l'intention d'un certain public. De nos jours, la traduction automatique rend des services inestimables dans le contexte l'assimilation, du moins pour certaines paires de langues et certains types de documents. Elle se révèle efficace quand le résultat de traduction doit êre obtenu rapidement et que la qualité de traduction n'a pas besoin d'être parfaite. Dans ce cas, l'utilisateur met l'emphase sur l'adéquation du sens de la traduction par rapport à la phrase source. La traduction automatique se révèle beaucoup moins adaptée aux scénarios de dissémination où la qualité de la traduction est importante car elle a un impact sur l'image de l'auteur. Pour atteindre la qualité de traduction nécessaire à un but de dissémination, il est très souvent nécessaire de faire appel à des traducteurs professionnels.

Au delà des deux usages précédemment cités, les récents développements technologiques (en particulier ceux du Web) ont fait émerger de nouvelles fonctions pour la traduction automatique. En 2004, John Hutchins définit deux fonctions distinctes : une fonction d'échange qui consiste à utiliser la traduction automatique comme interprète simultané de textes électroniques rédigés dans une langue étrangère (discussions instantanées, courriers électroniques, etc.) ; et une fonction d'accès à l'information en langue étrangère par l'interrogation des systèmes de TA pour la traduction de pages Web.

L'objectif des recherches en traduction automatique qui prétendait, à ses prémices, une traduction entièrement automatique et de haute qualité (désignée par le terme « *Fully Automatic High-Quality Translation* » en anglais), a été rapidement revu de façon plus modeste. De nos jours, au delà de sa fonction d'assimilation de l'information, la traduction automatique est de plus en plus envisagée comme un outil d'aide aux traducteurs professionnels.

1.3 Limites et enjeux de la traduction automatique

Alors que la demande potentielle est énorme, la traduction automatique reste aujourd'hui un marché à exploiter. Face à l'explosion de la diffusion de contenus multilingues, les enjeux sont à la fois économiques, géopolitiques et culturels.

La traduction automatique offre déjà de réels services et les systèmes grand public que l'on trouve sur le Web peuvent nous donner une idée sommaire du contenu d'une page. Cependant, même les logiciels actuels les plus perfectionnés ne peuvent produire des résultats de traduction équivalent à ceux d'une personne de langue maternelle ou possédant les compétences d'un traducteur professionnel. En effet, bien que le maniement de la langue soit tout naturel pour les êtres humains, sa modélisation pose d'immenses problèmes. La traduction demeure donc une activité linguistique et une tâche complexe, l'automatiser est un exercice extrêmement difficile.

Bien que la traduction automatique ait prouvé son efficacité pour traduire des textes de domaines spécifiques et restreints (la traduction de bulletins météorologistes obtient, par exemple, de très bons résultats), de nombreux facteurs expliquent que la traduction automatique de la langue générale ne fournisse pas toujours des résultats corrects : le langage naturel est subtil, équivoque et il peut exister de nombreuses traductions et interprétations différentes possibles de certains mots et phrases. Les termes polysémiques, l'homonymie, les structures grammaticales ambiguës, les ambiguités référentielles et autres problèmes grammaticaux représentent autant de sources d'erreurs pour un système de traduction automatique. Quelques exemples de traductions erronnées sont donnés ci-après :

Termes polysémiques « *La vedette accoste au port.* » est traduit par « *The star accosts with the port.* » par Systran[5]. En l'absence de contexte, « vedette » peut avoir le sens de « star » ou de « sentinelle » mais il a ici le sens de « petit bateau à moteur » traduit par « speedboat » ou « launch » en anglais ;

5. http ://www.systranet.com/fr/traduction

Homonymie « *Elle a perdu ses fils et ne peut donc plus coudre.* » est traduit par « *She lost her son and can no longer sew.* » par Google [6]. Ici, d'après le contexte, « fils » devrait prendre le sens de « brins longs et fins de matière textile » plutôt que celui de « personne descendante de sexe masculin » ;

Ambiguité référentielle « *Paul a trouvé la bague de sa femme. Il la lui rend.* ». Il est impossible pour un système de traduire correctement sans savoir si le pronom « la » réfère à la femme ou à la bague ;

Expressions floues « *Parler n'est-il pas toujours en un sens donner sa parole ?* » traduit « *To speak isn't always in a direction to give its word ?* » par Systran ;

Expressions idiomatiques « *Jean et lucien sont désormais à couteaux tirés.* » traduit « *Jean and Lucien are from now on with drawn knives.* » par Systran ;

Transposition « *Detroit has a bunch of run-down houses for sale.* » traduit « *Detroit a un groupe de maisons de course vers le bas à vendre.* » par Systran.

Si certains problèmes relatifs à la traduction automatique sont dus à une modélisation insuffisante des dynamiques linguistiques de la langue, la plupart d'entre eux proviennent d'un manque de considération du contexte sémantique et de la question de la « connaissance du monde ».

Comme le faisait remarquer Y. Bar-Hillel dès 1960 dans [Bar-Hillel 1960] : « *L'activité de traduction automatique nécessite le recours fréquent à des connaissances du domaine traité. Il n'était pas absurde a priori d'espérer qu'avec des moyens différents, ce recours puisse s'avérer être superflu. L'expérience dément cet espoir. Comme il est chimérique de vouloir représenter en machine toute la connaissance, disons d'une encyclopédie, la traduction entièrement automatique de haute qualité est impossible* ».

L'étape décisive consistera donc à passer de la représentation syntaxique actuelle à une représentation sémantique du texte. L'interprétation peut être vue comme la première étape de la compréhension qui reste, pour le moment, une faculté humaine. De ce fait, beaucoup de chercheurs s'accordent sur le fait que les systèmes de traduction de la langue générale n'évoluent plus, principalement car la traduction automatique contemporaine bute toujours sur le même problème : aucun système n'a de compréhension globale du monde.

Cependant, même si les logiciels informatiques de traduction ne sont pas en mesure de substituer complètement aux humains, ils peuvent apporter une aide réelle pour des usages centrés sur la compréhension de langues peu ou mal connues de l'utilisateur. Le succès de ces outils auprès des utilisateurs exprime un besoin de traduction rapide (voire instantannée) au détriment de la qualité qui peut rester grossière si la necessité est juste de comprendre le sens d'un texte. L'amélioration des performances et de la rapidité des processeurs des micro-ordinateurs, les progrès réalisés dans les domaines de la linguistique et de l'informatique permettent aujourd'hui d'utiliser sur des postes de travail standard des logiciels de bonne qualité qui permettent d'obtenir en « premier jet » l'accès à un texte en langue étrangère.

Même si, dans leur état actuel, les outils de traduction automatique offrent déja de grands services et ont gagné leur place auprès des technologies d'aujourd'hui, ils ont

6. http ://translate.google.fr

aussi leurs limites et l'intervention humaine, même réduite, restera toujours nécessaire pour comprendre et retranscrire toutes les subtilités de la langue.

1.4 Méthodologies pour la traduction automatique

Les tendances actuelles de la recherche en traduction automatique répartissent les chercheurs du domaine entre deux approches computationnelles dominantes. Une première utilise des méthodes expertes et vise à formaliser toutes les connaissances nécessaires à la traduction (sous la forme, par exemple, de dictionnaires et grammaires). La seconde, basée sur des méthodes empiriques fait en sorte, quand à elle, que toute connaissance linguistique soit acquise de manière empirique et automatique à partir de corpus. Parmi elles, la traduction automatique statistique utilise des modèles stochastiques qui sont construits à partir du traitement automatique d'une grande quantité de données.

Historiquement, la première approche utilisée pour traduire des textes fut la méthodologie experte, basée sur des règles linguistiques. L'ensemble des règles définit les possibilités d'association de mots, selon leurs catégories lexicales, et permet de modéliser la structure d'une phrase donnée. Celle-ci nécessite beaucoup de travail de la part des linguistes pour définir le vocabulaire et la grammaire. Cette méthode fournit des résultats étonnants mais elle montre tout de même rapidement ses limites du fait qu'il soit sans cesse nécessaire d'avoir recours à des experts bilingues. Par ailleurs, un des inconvénients des modèles construits sur une approche structurelle est la conception des grammaires, c'est-à-dire la définition des règles. Celles-ci doivent, en effet, être suffisamment robustes pour prendre en compte tous les phénomènes syntaxiques d'une langue.

Depuis, d'autres types d'approches se sont développés : l'approche à base d'exemples et l'approche statistique qui font en sorte que toute connaissance linguistique soit acquise ou codée de manière empirique et automatique à partir de grandes quantités de textes (des distributions de probabilités pouvant ici remplacer les règles). Un des avantages des approches statistiques est leur indépendance vis-à-vis de l'expertise des linguistes, même si l'intervention humaine reste tout de même nécessaire pour générer les exemples bilingues alignés sur lesquels apprennent les systèmes. Avec l'apparition progressive de grandes quantités de données multilingues ainsi que le développement et la maintenance de boîtes à outils, cette approche est maintenant dominante et fera l'objet de la suite de ce manuscrit.

2 Modèles de référence pour la traduction probabiliste

2.1 Enoncé du problème

La modélisation de la traduction automatique statistique repose sur la théorie mathématique de distribution et d'estimation probabiliste développée en 1990 par Frederick Jelinek de la société IBM au Thomas J. Watson Research Center. Ces bases

conceptuelles ont été introduites dans [Brown et al. 1990].

L'hypothèse initiale est que toute phrase dans une langue est une traduction possible d'une phrase dans une autre langue. Si on traduit depuis une langue source s vers une langue cible t, le but est de trouver la phrase cible t la plus appropriée pour traduire la phrase source s. Pour chaque paire de phrase possible (s, t), on attribue une probabilité $P(t|s)$ qui peut être interprétée comme la probabilité que « un traducteur » produise t dans la langue cible, lorsque la phrase s a été énoncée dans une langue source, ou autrement dit, la probabilité que la traduction de s soit t. On espère, par conséquent, qu'une paire comme *(Le petit chat est noir, The french country is famous for its red wine)* obtienne une probabilité faible tandis que *(Le petit chat est noir, The little cat is black)* ait une probabilité élevée. Dans la traduction automatique statistique, les modèles probabilistes sont utilisés pour trouver la meilleure traduction possible t^* d'une phrase source donnée s, parmi toutes les traductions t possibles dans la langue cible. On veut donc trouver le t^* qui maximise $P(t|s)$, ceci s'exprime par l'équation I.1. La recherche de cette meilleure traduction est appelée le décodage.

$$t^* = argmax_t P(t|s) \tag{I.1}$$

Le but des méthodes probabilistes pour la traduction automatique est de construire des systèmes de traduction exploitant l'équation I.1. Il s'agit alors d'appliquer des méthodes d'apprentissage statistiques afin d'entraîner le système avec des millions de mots de textes, dont des textes monolingues en langue source et des textes alignés composés d'exemples de traduction entre les deux langues.

2.2 Notion de corpus

Avec les systèmes de traduction statistique, les hypothèses de traductions sont générées sur la base de modèles statistiques dont les paramètres sont estimés à partir de l'analyse d'une grande quantité de données d'apprentissage monolingue et bilingue : les corpus.

Un corpus est un ensemble de documents regroupés dans un but précis. La notion de corpus est utilisée dans plusieurs domaines : études littéraires, linguistiques, scientifiques, etc. Les corpus sont indispensables en traitement automatique du langage naturel : ils permettent en effet d'extraire un ensemble d'informations utiles pour des traitements statistiques. Les corpus dont nous parlerons ici sont, plus précisément, de grandes collections de textes représentatifs d'une langue, d'une paire de langues, voire d'un domaine.

En traduction automatique statistique, les corpus sont composés d'exemples de traduction entre les deux langues, ou plus précisément d'un ensemble de phrases en langue source auxquelles sont assimilées une ou plusieurs traductions en langue cible. C'est ce que l'on appelle des corpus parallèles.

La création d'un système de traduction probabiliste nécessite des corpus pour l'entraînement, le développement et l'évaluation du système. Le corpus d'apprentissage a pour but d'entraîner et de construire les modèles à l'aide de méthodes d'apprentissage statistique, le corpus de développement sert à « ajuster » et améliorer les modèles appris alors que le corpus de test permet de vérifier et tester la qualité du modèle appris.

2.3 Equation fondamentale

Les premiers modèles pour la traduction automatique statistique ont été présentés dans [Brown et al. 1991] et [Brown et al. 1993] par les chercheurs d'IBM du Thomas J. Watson Research Center. Dans cette approche, la traduction est vue comme le problème de trouver le t le plus probable sachant s, étant donnée la décomposition avec la règle de Bayes :

$$P(t|s) = \frac{P(s|t) \times P(t)}{P(s)}$$

Dans cette formule, le dénominateur $P(s)$ est indépendant de la phrase cible t cherchée, ce qui signifie que la recherche de la meilleure traduction t^* peut se réduire à l'équation I.2.

$$t^* = argmax_t P(t|s) = argmax_t P(s \mid t) \times P(t) \tag{I.2}$$

Dans cette formule, $P(t)$ est appelé le modèle de langage de la langue cible et $P(s|t)$ est appelé le modèle de traduction. Ces deux modèles sont appris empiriquement à partir de corpus. $P(t)$ est la probabilité de la phrase t dans la langue cible et $P(s|t)$ est la probabilité que la phrase source s soit une traduction de la phrase cible t. La phrase t^* retenue pour la traduction de la phrase s sera celle qui maximise les deux modèles probabilistes de l'équation I.2.

Modèle de langage

Comme vu dans l'équation I.2, $P(t)$ représente le modèle de langage, c'est-à-dire la composante du système de traduction qui est en charge d'introduire les contraintes imposées par la langue cible.

La probabilité $P(t)$ estime *a priori* la vraisemblance de la séquence de mots ou phrase t. Ainsi, plus la séquence de mots t sera conforme au modèle de langage, plus sa probabilité $P(t)$ sera élevée.

Il existe différentes techniques pour le calcul des probabilités d'un modèle de langage. Celles-ci sont généralement apprises sur des corpus d'apprentissage monolingue censés représenter au mieux la langue cible à modéliser. Dans ce manuscrit, nous nous intéresseront aux modèles de langage à base de n-grammes qui sont actuellement les modèles les plus utilisés dans le domaine de la traduction automatique probabiliste. Un n-gramme est une sous-séquence de n mots continus, construite à partir d'une séquence de mots donnés. La modélisation des modèles de langage à base de n-grammes repose sur l'hypothèse que la probabilité conditionnelle d'apparition d'un mot ne dépend que des $n-1$ mots qui le précèdent. En pratique, les modèles 3-grammes (dits trigrammes), 4-grammes et 5-grammes, s'avèrent les plus performants et les plus couramment utilisés pour modéliser les différentes langues européennes.

La probabilité $P(t)$ d'une séquence de mots t, donnée par un modèle de langage statistique n-gramme est calculée comme suit :

$$P(t) = \prod_{i=1}^{k} P\left(w_i|w_{i-1}w_{i-2}\cdots w_{i-n+1}\right)$$

avec $t = w_1 w_2 w_3 \cdots w_k$ où w_i sont des mots, n la taille maximale des n-grammes utilisés dans le modèle de langage et $P\left(w_i | w_{i-1} w_{i-2} \cdots w_{i-n+1}\right)$ la probabilité du mot w_i sachant les $n-1$ mots qui le précèdent.

Modèle de traduction

$P(s|t)$ est la probabilité que la phrase s soit une traduction de la phrase t. On apprend $P(s|t)$ à partir d'un corpus bilingue aligné en phrases (à une phrase en langue source correspond une phrase en langue cible). Etant donné que les données du corpus ne sont généralement pas suffisantes pour apprendre directement $P(s|t)$, on décompose les phrases s et t en unités plus petites. On a alors $s = s_1 s_2 \cdots s_N$ et $t = t_1 t_2 \cdots t_M$ où N et M sont les nombres de subdivisions des phrases. Les unités de s sont ensuite mis en correspondance avec les unités de t selon une technique d'alignement. Le modèle de traduction nécessite donc d'apprendre des alignements entre les unités du corpus parallèle. Pour cela, plusieurs approches existent et il est nécessaire d'introduire la notion d'alignement.

Notion d'alignement

La quasi-totalité des modèles de traduction $P(s|t)$ introduisent une variable cachée A, appelée alignement, qui décrit une correspondance entre les unités d'une phrase et ceux de sa traduction.

Un modèle statistique de traduction évalue, par la quantité $P(s|t)$, la probabilité que la phrase $s = s_1 s_2 \cdots s_N$ soit une traduction de la phrase $t = t_1 t_2 \cdots t_M$ où les t_j et s_i sont des unités des phrases t et s. En pratique ces unités sont soit des mots, soit des groupes de mots. La figure I.3 montre un exemple d'alignement en mots et la figure I.4 un exemple d'alignement en groupes de mots.

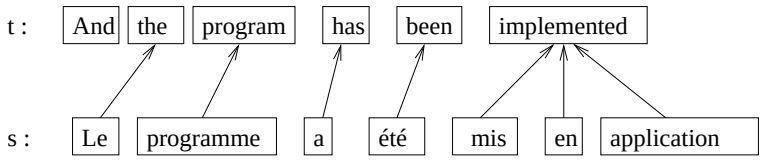


FIGURE I.3 – Exemple d'alignement en mots pour la traduction

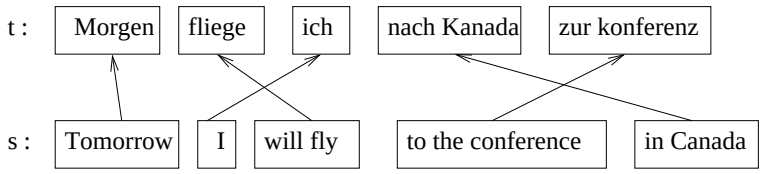


FIGURE I.4 – Exemple d'alignement en segments pour la traduction

La probabilité de traduction est estimée par la somme des alignements possibles entre les unités de s et celles de t, on considère donc $P(s|t) = \sum_{a \in A} P(s, a|t)$, avec a

un alignement possible entre la phrase source et la phrase cible. La cardinalité de A est néanmoins trop grande pour être calculée directement car le nombre d'alignements croît très rapidement avec le nombre d'unités des phrases : pour une phrase source de N mots et une phrase cible de M mots, on a $(N+1)^M$ alignements possibles entre les mots. En pratique, on approxime donc la somme de tous les alignements possibles par la probabilité de l'alignement le plus probable. Pour trouver cet alignement de probabilité maximale, on utilise l'algorithme de Viterbi.

Par la suite, nous considèrerons deux techniques d'alignement : l'alignement en mots (dit « word-based ») où les phrases sont décomposées en mots, et l'alignement en segments (dit « phrase-based ») où l'unité de division de la phrase est le groupe de mots.

2.4 Modèles de traduction à base de mots

Dans les modèles de traduction à base de mots, les mots sont les unités fondamentales de traduction. Le but est donc d'aligner les corpus paralèlles au niveau des mots. Les modèles théoriques pour l'alignement en mots, que je présente ci-après, ont étés proposés par IBM.

Les modèles IBM

Les articles de référence des méthodes de traduction probabilistes à base de mots d'IBM font mention de 5 modèles de traduction dont le but est d'évaluer la probabilité $P(t|s)$. Le premier modèle d'IBM décrit dans [Brown et al. 1990] considère la distribution des mots comme uniforme, donc tous les alignements comme équiprobables (il ne prend pas en compte l'ordre des mots). Il faut attendre le deuxième modèle d'IBM, décrit dans [Brown et al. 1991], pour que l'ordre des mots soit pris en compte. Celui-ci intègre en effet un modèle de distorsion, autrement appelé modèle de ré-ordonnancement, qui représente la distance entre la position d'un mot de la phrase source s et celle d'un mot de la phrase cible t qu'il a produit. D'autre part, ce modèle permet l'alignement à un mot spécial appelé *null* utilisé lorsqu'un ou plusieurs mots d'une phrase n'ont pas de correspondant dans l'autre phrase (formellement, il y a un mot *null* dans chacune des langues). La probabilité $P(t|s)$ est calculée à partir du corpus parallèle préalablement aligné avec un apprentissage par maximum de vraisemblance. La vraisemblance est maximisée avec l'algorithme *expectation-maximisation* ou EM [Dempster et al. 1977].

Le troisième modèle d'IBM introduit la notion de fertilité qui autorise un mot source à générer plusieurs mots cibles. C'est le modèle le plus utilisé dans les systèmes de traduction probabilistes car il présente le meilleur rapport complexité/efficacité. Le modèle de traduction $P(s|t)$, dont la formule est donnée dans l'équation I.3, dépend alors de 4 paramètres :

- $P(\hat{n}_a(s)|s)$ est le modèle de fertilité où $\hat{n}_a(s)$ est le nombre de mots de t alignés avec s dans l'alignement a ;
- $P(t|\hat{s}_a(t))$ est le dictionnaire des mots où $\hat{s}_a(t)$ est le mot de s aligné avec t dans l'alignement a ;

– $P_{distorsion}(t, a, s)$ est le modèle de distorsion qui est calculé avec s_i, la position d'un mot de s, et t_j, la position de sa traduction dans t, et N, M, respectivement, le nombre de mots des phrases source et cible.

$$P(t, a|s) = \prod_{t \in T} P(t|\hat{s}_a(t)) \times \prod_{s \in S} P(\hat{n}_a(s)|s) \times P_{distorsion}(t, a, s) \tag{I.3}$$

Les modèles 4 et 5 d'IBM sont identiques au modèle 3 (*cf* équationI.3) mais utilisent une modélisation plus complexe de ré-ordonnancement.

Limitations de l'alignement en mots

Cette méthode d'alignement en mots trouve ses limites au niveau des phrases dont la traduction de certains mots dépend de la traduction d'autres mots de la phrase. Le problème est dû au fait que les modèles à base de mots partent du principe que les mots sont indépendants les uns des autres. La notion de fertilité, en effet, autorise un mot source à être aligné avec plusieurs mots cibles (traductions m-à-1) mais, comme le montre la figure I.5, n'autorise pas que plusieurs mots cibles soient alignés avec un seul et même mot source (traductions 1-à-n). Dans la traduction mot à mot, les modèles IBM ne sont capables de générer que des correspondances m-à-1. Or, pour prendre en compte toute la complexité du langage, il est nécessaire de pouvoir générer des correspondances n-à-m. Pour cela, il est indispensable de considérer, non plus les mots, mais des suites ou séquences de mots.

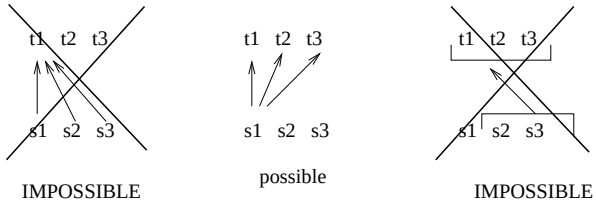


FIGURE I.5 – Alignements en mots avec le modèle IBM

2.5 Modèles de traduction à base de segments

Dans la suite de ce rapport, on désignera par « segment » une séquence de mots contigus qui n'est pas forcément un groupe syntagmatique au sens linguistique du terme.

Dans [Och et Ney 2002], les auteurs proposent de pallier aux difficultés que rencontre l'alignement en mots des modèles IBM, en introduisant la notion d'alignement de segments de phrases. Le but est de pouvoir construire des alignements n-à-m. Le segment devient alors l'unité sur laquelle se fonde la traduction. On peut alors capturer des dépendances entre les mots. Cette notion est reprise dans [Tomas et Casacuberta 2001] puis élaborée dans [Koehn et al. 2003] sous le terme de *table de traduction*. Une

table de traduction contient tous les alignements en segments ainsi que leurs probabilités. Les auteurs de ce dernier article comparent trois méthodes pour construire la table de traduction. Il en résulte que la méthode la plus efficace est celle qui utilise la cohérence des blocs de l'alignement en mots présentée par [Och et al. 1999]. Très vite, les modèles de traduction basés sur les segments se sont révélés plus performants que ceux basés sur les mots.

Les principaux composants de l'équation des modèles de traduction basés sur les segments, donnée dans la formule I.4, ci-après, sont alors :

- le modèle de langage : $P(t)$;
- le modèle de traduction : $P(s,t)$;
- le modèle de distorsion ou ré-ordonnancement : $\Omega(s|t)$.

$$P(t|s) = P(t) \times P(s,t) \times \Omega(s|t) \tag{I.4}$$

Alignement en segments

Plusieurs heuristiques ont alors été proposées pour effectuer l'alignement des phrases en segments et créer la table de traduction. La plus usitée est celle qui consiste à aligner le corpus en segments de phrases à partir de l'alignement en mots du corpus [Och et al. 1999]. Les principales étapes de l'alignement en segments d'un corpus sont les suivantes :

- on crée un corpus bi-directionnel aligné en mots en effectuant un alignement, selon un modèle à base de mots, dans le sens phrase source vers phrase cible puis dans le sens inverse phrase cible vers phrase source ;
- on considère l'intersection des deux alignements en mots puis l'on extrait tous les segments cohérents avec l'alignement en mots. Cela donne un fort taux de précision sur les alignements possibles mais un rappel relativement faible ;
- on essaie d'élargir l'intersection pour augmenter le rappel sans perdre en précision. Il y a plusieurs statégies possibles qui dépendent de la taille du corpus et de la paire de langue ;
- après avoir collecté les paires alignées de segments, on estime leur probabilité avec leur fréquence relative afin de construire la table de traduction. On attribue ensuite les probabilités $P(\bar{s}|\bar{t}) = \frac{count(\bar{s},\bar{t})}{\sum_s count(\bar{s},\bar{t})}$ et $P(\bar{t}|\bar{s}) = \frac{count(\bar{s},\bar{t})}{\sum_t count(\bar{s},\bar{t})}$ à chaque bi-segment en effectuant le comptage sur le corpus parallèle. $\bar{s}$ et $\bar{t}$ désignent respectivement un segment source et un segment cible.

Lors de la traduction, la phrase source s est alors décomposée en segments. Tous ces segments sont traduits en langue cible à l'aide du modèle de traduction et les segments cibles sont ensuite ré-ordonnancés avec le modèle de distorsion.

Une des limites des modèles basés sur les segments concerne leur lacune à bien gérer les segments non contigus (une extension a été proposée par [Simard et al. 2005]). D'autre part, il est à noter que l'espace nécessaire pour stocker la table de traduction et le nombre d'hypothèses à explorer augmente considérablement avec la taille du corpus d'apprentissage. Il existe cependant des travaux visant à réduire la taille de la table de traduction par des méthodes de filtrage [Johnson et al. 2007].

2.6 Modèle log-linéaire

Tels qu'ils ont évolué, les modèles probabilistes actuels peuvent inclure un modèle de langage, un modèle de traduction, un modèle de ré-ordonnancement, etc. Pour combiner ces différents modèles, [Och et Ney 2002] ont proposé d'utiliser un modèle linéaire discriminant, avec la log-probabilité comme fonction caractéristique.

D'une façon générale, le modèle log-linéaire usuel a pour fonction de prédire la valeur d'une variable attendue à partir d'un ensemble de variables descriptives non nécessairement indépendantes. La méthode est aujourd'hui largement appliquée dans divers domaines.

Le principal avantage de ce modèle est qu'il permet l'ajout aisé de « nouvelles » variables descriptives. Néanmoins, en contrepartie, un grand nombre de variables descriptives est susceptible de rendre difficile l'interprétation des résultats.

Dans le cas de la traduction automatique probabiliste, un modèle de combinaison log-linéaire est utilisé à la place d'une simple combinaison linéaire car les différentes valeurs des probabilités de $P(\bar{t}|\bar{s})$ diffèrent souvent en ordre de grandeur. Dans un tel cas, la combinaison log-linéaire est meilleure que la combinaison linéaire car celle-ci présuppose que chaque fonction fournit une quantité d'information proportionnelle à sa probabilité correspondante. Le modèle de traduction initial tel que décrit par IBM a donc, par la suite, évolué en un modèle log-linéaire. Ce modèle est pour la première fois utilisé en traduction automatique dans [Och et Ney 2002]. Cet article décrit la manière dont le modèle combine différentes composantes (modèle de traduction, modèle de langage, etc) en utilisant l'estimation des paramètres par maximisation de vraisemblance.

Selon [Oepen et al. 2007], un modèle log-linéaire est définit par (a) un ensemble de caractéristiques spécifiques qui décrivent les propriétés des données et (b) un ensemble associé de poids qui déterminent la contribution de chaque caractéristique. Le principe de ce modèle est donc une combinaison log-linéaire de plusieurs composantes h_i pondérées par des poids λ_i. Le facteur de pondération λ_i est utile pour introduire l'importance que l'on confère à une fonction de caractéristique h_i.

On estime la probabilité $P(\bar{t}|\bar{s})$ directement à partir d'un corpus bilingue aligné à l'aide d'un modèle log-linéaire exprimé comme dans l'équation I.5 où $h_m(\bar{t}, h, \bar{s})$ sont les fonctions caractéristiques, qui peuvent être vues comme des variables quantitatives, et λ_m sont les poids de ces fonctions.

$$P(\bar{t}|\bar{s}) = \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\bar{t}, h, \bar{s})\right) \tag{I.5}$$

La meilleure traduction, t^*, est donnée par l'équation :

$$t^* = argmax_t P(\bar{t}|\bar{s})$$

On remarquera que le modèle de référence, vu dans l'équation I.2 de la partie 2.3 page 13, est un cas particulier du modèle log-linéaire où $h_1 = logP(s|t)$, $h_2 = logP(t)$ et $\lambda_1 = \lambda_2 = 1$, ce qui donne :

$$argmax_t P(\bar{t}|\bar{s}) = argmax_t \left(\exp\left(logP(\bar{s}|\bar{t}) + logP(\bar{t})\right)\right) = argmax_t \left(P(\bar{t}) \times P(\bar{s}|\bar{t})\right)$$

Les avantages du modèle log-linéaire pour la traduction automatique probabiliste résident dans la possibilité de contrôler différents aspects de la qualité de la traduction, moduler l'importance que l'on confère à chaque caractéristique de la traduction et ajouter facilement et directement de nouvelles fonctions caractéristiques. Les modèles log-linéaires représentent, de nos jours, l'état de l'art de la traduction automatique probabiliste.

Les fonctions caractéristiques

Les fonctions caractéristiques sont utilisées dans les approches de traduction automatique pour sélectionner et classer les traductions candidates. La combinaison linéaire des fonctions caractéristiques, developpée dans un cadre de traitements statistiques, présente un cadre flexible pour la modélisation discriminante et permet de combiner les différentes sources d'information d'un même modèle. Le modèle log-linéaire repose sur un ensemble de fonctions caractéristiques dont chacune donne des informations sur un aspect des caractéristiques d'une bonne traduction. En général, on utilise, au moins, les principaux modèles implémentés par IBM tels que le modèle de traduction, le modèle de distorsion, et la pénalité sur les mots. On peut cependant définir un modèle log-linéaire plus riche en y ajoutant des composants comme les cinq modèles de ré-évaluation de [Mauser et al. 2006], ou la kyrielle de fonctions de [Och et al. 2004]. Toute fonction aidant à produire une traduction correcte peut être incluse, sans autre justification théorique. Un système de traduction compte en général entre cinq et une quinzaine de ces fonctions caractéristiques.

Les poids du modèle

En pratique, il est souvent bénéfique de pondérer les différentes sources d'information. Dans le modèle log-linéaire, la contribution de chaque composant est spécifiée par un poids qui va déterminer son importance. Dans l'équation I.5, la contribution de chacune des fonctions prédéfinies h_m est pondérée par un facteur λ_m à estimer. Ces λ_m constituent les paramètres du modèle.

Il a été montré qu'affecter des valeurs pertinentes à ces poids améliore de façon significative la qualité de la traduction. Les valeurs des poids doivent être, en effet, en adéquation avec le corpus traité et la paire de langue utilisée. Le modèle de pénalité sur les mots, par exemple, favorise les traductions longues s'il est pondéré par une valeur négative alors qu'un poids positif favorise les traductions courtes.

En pratique, on utilise un corpus de développement pour optimiser les valeurs des paramètres. Tel que décrit dans [Koehn et al. 2003], l'ajustement des paramètres du modèle à l'aide d'un corpus de développement est une technique qui permet d'améliorer considérablement la performance des modèles log-linéaires. Les poids y sont optimisés sur des données de développement différentes des données d'apprentissage et réprésentatives des données de test, sur lesquelles va être évalué le système final.

Optimisation des poids du modèle

L'optimisation des poids du modèle log-linéaire est une étape cruciale pour adapter à des données et de façon optimale un système de traduction statistique reposant sur un modèle linéaire discriminant. Le problème qui émerge est : comment optimiser les poids des fonctions caractéristiques ? En d'autres termes, comment trouver les poids qui offrent la meilleure qualité de traduction.

La fonction log-linéaire de décodage étant non dérivable, les algorithmes efficaces d'optimisation par descente de gradient ne peuvent donc pas être utilisés.

L'algorithme *Minimum Error Rate Training* La solution courante est d'utiliser la procédure introduite dans [Och 2003], *Minimum Error Rate Training* (MERT). Cette approche cherche les poids qui minimisent une mesure d'erreur donnée, pour un corpus de développement donné. Les valeurs des paramètres λ_m sont alors choisies de façon à maximiser la qualité des traductions produites sur ce corpus de développement. On cherche alors le vecteur de poids des paramètres qui minimise la distance entre la traduction du texte source donnée par le système et la traduction cible du corpus parallèle (appelée dans ce cas « la traduction de référence ») selon une métrique choisie. On retient la combinaison des poids qui correspond à la performance optimale du système de traduction mesurée sur le corpus de développement. Si le corpus de développement est représentatif du corpus de test, on peut espérer une amélioration des résultats sur le corpus de test seront également améliorés.

Cet algorithme, représenté dans le schéma I.6, optimise les poids et permet au décodeur de produire des traductions de meilleure qualité (selon une métrique Err et une ou plusieurs références ref) sur un corpus parallèle de développement.

$$t^* = argmax_t P_{\lambda^*}(\bar{s}, \bar{t}) \;\; avec \;\; \lambda^* = argmin_\lambda Err\,(t^*(\lambda, ref)) \tag{I.6}$$

Le paramètre Err de l'équation I.6 représente une méthode de mesure de performance automatique que l'on choisira (par exemple BLEU, NIST, WER, etc.). Elle sera utilisée pour calculer la distance entre la traduction de référence et la traduction du système.

Une méthode initiale consisterait à rechercher les valeurs optimales des paramètres parmi toutes les combinaisons des paramètres possibles mais il est impossible d'explorer de façon exhaustive tout l'espace de recherche car celui-ci croît de façon excessive avec le nombre de poids à optimiser.

En tenant compte de cette contrainte, l'approche utilise les meilleures traductions alternatives (pour chaque phrase source fournie) comme une approximation de la sortie du décodeur, ce qui permet une convergence plus rapide du processus d'optimisation. La métrique d'évaluation automatique des traductions utilisée dans MERT est le score BLEU (mais celui-ci peut être remplacé par une quelconque autre métrique automatique). L'algorithme d'optimisation des poids d'Och, cherche à mettre à jour les poids des fonctions caractéristiques de façon à améliorer le score d'évaluation. Comme décrit dans le schéma I.6, pour chaque phrase source du corpus de développement, l'algorithme MERT consiste à décoder cette phrase source avec les poids actuels ; générer une liste

des N meilleures traductions ; optimiser la valeur des λ_i ; re-décoder les N meilleures traductions avec ces nouveaux paramètres, ré-optimiser ceux-ci, etc. On itère un certain nombre de fois l'ajustement des poids des fonctions jusqu'à observer une convergence.

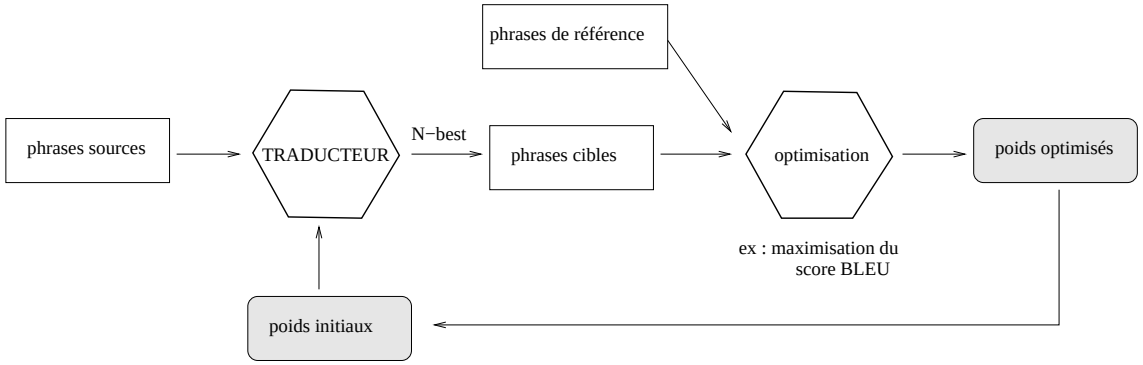


FIGURE I.6 – Apprentissage des poids des fonctions caractéristiques

Il est nécessaire de préciser qu'une augmentation de la performance sur le corpus de développement n'entraîne pas forcément une augmentation de la performance sur le corpus de test. Deux vecteurs de poids donnant approximativement le même score BLEU sur l'ensemble de développement, pourront donner deux scores BLEU différents sur l'ensemble de test. Jusqu'à présent, aucune méthode permettant de prédire, de façon sûre, l'augmentation ou non de la performance sur le corpus de test n'a été présentée. Pour ces raisons, entre autres, il arrive que l'algorithme MERT ne trouve pas un bon maximum sur l'ensemble de développement ou que le maximum trouvé ne conduise pas à une amélioration sur l'ensemble de test.

Bien que très largement utilisée dans la communauté de la traduction automatique probabiliste, la méthode *Minimum Error Rate Training* présente plusieurs inconvénients, dont, par exemple, le fait de ne pas supporter l'utilisation d'un trop grand nombre de fonctions de caractéristiques. En réponse à ces faiblesses constatées, plusieurs alternatives à la méthode ont été proposées. Parmi celles-ci, MIRA (pour *Margin Infused Relaxed Algorithm* en anglais) est un algorithme d'apprentissage « à la volée » (les poids sont actualisés à la suite du décodage de chaque phrase du corpus de développement) qui supporte l'utilisation de plusieurs milliers, voire millions, de fonctions de caractéristiques [Chiang 2012, Hasler et al. 2011].

Apprentissage d'un modèle log-linéaire

Le processus d'apprentissage d'un système de traduction basé sur l'approche log-linéaire est présenté dans la figure I.7. Le système est constitué de plusieurs modèles $h_n(\bar{s}, \bar{t})$ appris à partir d'un corpus d'apprentissage sur lequel on extrait, entre autres, des probabilités de traduction entre des segments en langue source et en langue cible. Ces différents modèles appris sont ensuite affectés d'un poids λ_n et combinés dans un système log-linéaire qui servira de fonction de décision lors du décodage d'un corpus de test.

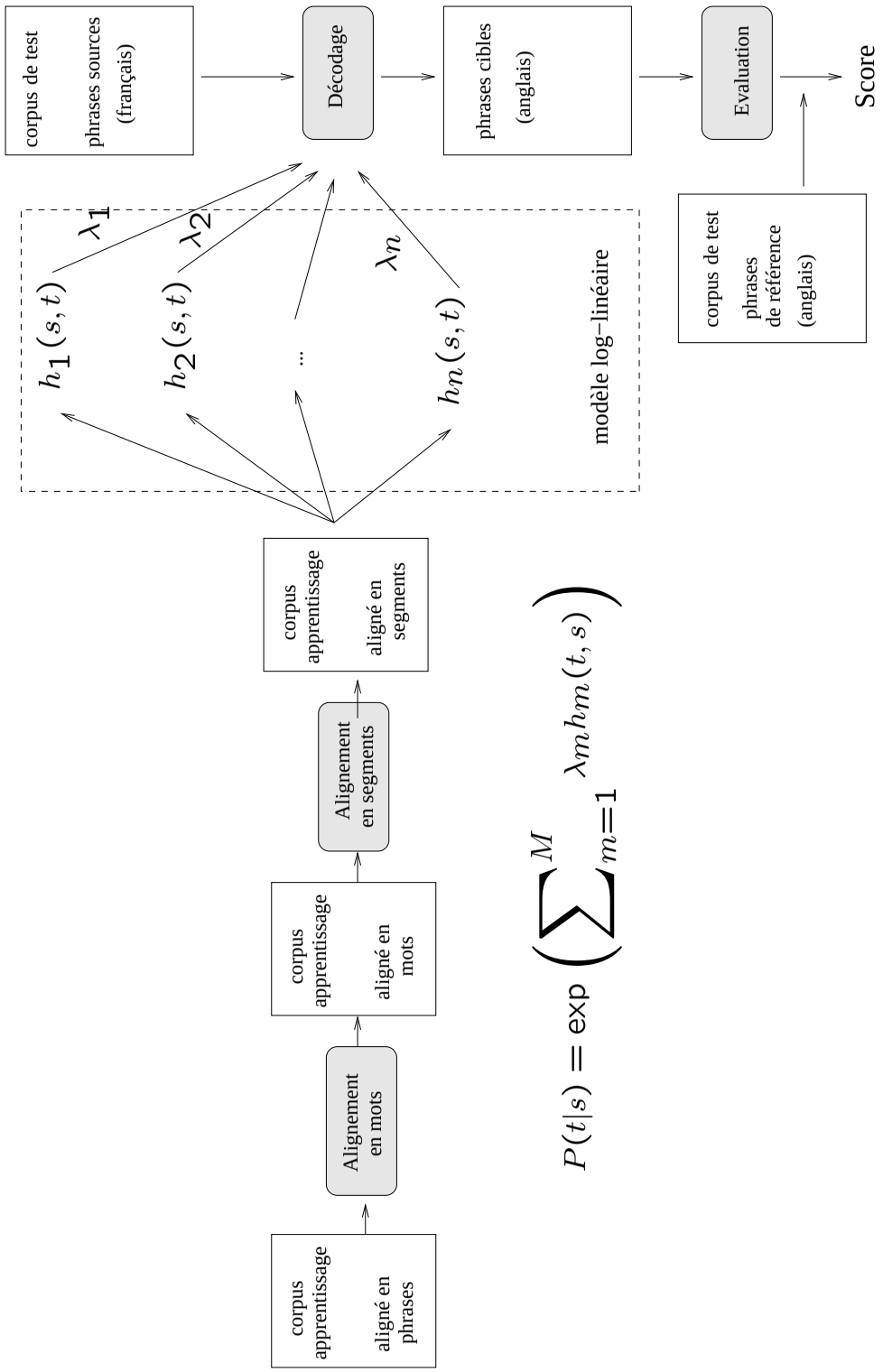


FIGURE I.7 – Etapes de création d'un système de traduction automatique statistique à base de segments

Alignement en mots du corpus d'apprentissage On aligne en mots chaque phrase du corpus d'apprentissage avec la technique vue dans la section 2.4 page 15 et présentée dans [Och et Ney 2003]. Pour cela on effectue un alignement de la phrase cible vers la phrase source puis dans la direction inverse. On obtient un alignement lexical bidirectionnel.

Alignement en segments du corpus d'apprentissage A partir de l'alignement bidirectionnel, on extrait des paires de séquences $(\bar{s}, \bar{t})$ selon une heuristique visant à trouver la réunion puis l'intersection adéquate en évaluant le taux de rappel et le taux de précision. On obtient alors une table de bi-segments $(\bar{s}, \bar{t})$.

Assignation de statistiques de traduction aux paires de séquences On associe, à chaque bi-segment de la table, une statistique de traduction calculée, dans les deux sens de traduction ($T(\bar{s}/\bar{t})$ et $T(\bar{t}/\bar{s})$), selon des fonctions de compte basées sur la fréquence relative.

Entraînement des différents modèles La table de traduction créée à l'étape précédente va servir à estimer les fonctions caractéristiques dépendantes de s et t dont entre autres le modèle de traduction, le modèle de langage, le modèle de distorsion, le modèle de pénalité sur les mots, etc.

Affectation des poids du modèle Le système final utilise les différentes fonctions caractéristiques apprises sur la table de traduction et sur le corpus d'apprentissage monolingue. Le modèle log-linéaire suppose d'affecter un poids à chacune de ces fonctions de caractéristique. Ce poids peut être affecté de façon arbitraire ou être ajusté à l'aide d'une technique d'optimisation que nous verrons dans la section suivante.

Décodage Le principe du décodage est de trouver, pour chaque phrase à traduire s, la traduction t^* qui maximise la probabilité du modèle log-linéaire $P(t|s)$. Lors du décodage, l'espace de recherche est extrèmement grand, c'est un problème NP-complet. Il est donc nécessaire de le réduire avec des heuristiques comme l'algorithme Beam Search qui élimine les hypothèses faibles le plus tôt possible.

Evaluation du système Généralement, en dernier lieu, on évalue la qualité du système créé avec une ou plusieurs techniques d'évaluation vues dans la section 3.

3 Evaluation de la qualité d'une traduction automatique

Une fois qu'une traduction est réalisée par un système automatique, il s'agit d'en évaluer sa « qualité ». En effet, une même phrase source : *« I carry out a try »*, peut

donner lieu à différentes traductions[7] : *« Je procéder à un essai »* ; *« Je réalise un essai »* ; *« J'effectue un essai »* ; *« J'emporte un essai »* ; *« J'exécute un essaie »*. Le but de l'évaluation de la traduction est d'attribuer une note de qualité à une hypothèse de traduction donnée. Il existe plusieurs approches pour évaluer la qualité de traductions.

3.1 Jugement humain

Lors d'une campagne d'évaluation humaine de traductions, un ou des annotateurs sont invités à juger les traductions d'un système en fonction de critère(s) de qualité donné(s) s'appliquant, par exemple, au texte, à la phrase ou aux segments des phrases. Les jugements humains sont communément effectués à l'échelle de la phrase dont la qualité de traduction est évaluée selon deux aspects : la fluidité et l'adéquation. La fluidité évalue à quel point la phrase traduite respecte la syntaxe de la langue de destination. Plus un texte se lit aisément, meilleure est sa fluidité. D'un autre côté, l'adéquation d'une traduction concerne la conservation du sens de la phrase source, le but étant de « s'attacher au texte de départ tout en respectant la destination (*nda* : le contexte d'usage) de sa traduction » (Translation Journal, October 2001).

Ces jugements qualitatifs sont très souvent traduits en termes de variables ordinales (très bonne, bonne, médiocre, etc.). L'évaluation humaine de la traduction est cependant une tâche très difficile. Plusieurs campagnes d'évaluation ont montré que l'accord entre annotateurs est généralement très faible [Callison-Burch et al. 2007, Callison-Burch et al. 2008, Callison-Burch et al. 2009]. Selon [Specia et al. 2010], une des explications à cet absence d'accord est la mauvaise spécification de la tâche : il est difficile, dans l'absolu, de définir ce qu'est une « bonne » traduction. Ce constat a fait récemment émerger un nouveau protocole d'évaluation. La qualité d'une traduction n'est alors plus évaluée dans l'absolu mais comparativement à d'autres traductions de la même phrase source. L'annotateur a, par exemple, pour tâche d'ordonnancer une liste de plusieurs hypothèses de traduction. Cette manière d'évaluer la traduction permet des jugements plus fiables avec un accord inter-annotateurs plus élevé [Callison-Burch et al. 2007, Callison-Burch et al. 2008, Callison-Burch et al. 2009].

Une telle évaluation subjective requiert une coûteuse intervention humaine et est, par ailleurs, sujette aux problèmes de non-reproductibilité et de variabilité inter annotateur. Bien souvent, pour des raisons de coût et de rapidité, on souhaite pouvoir évaluer la qualité d'une traduction sans recourir aux jugements humains. A cette fin, plusieurs mesures automatiques ont été développées au fil des années. Ces mesures sont censées être étroitement corrélées avec les scores que produirait une évaluation humaine.

3.2 Evaluation automatique

Par opposition à l'évaluation humaine, nous appellerons évaluation automatique un protocole ne nécessitant aucune intervention humaine au moment de l'évaluation. Bien

7. Les traductions données en exemple proviennent de systèmes de traduction automatiques disponibles gratuitement sur Internet.

qu'automatiques, ces systèmes d'évaluation ont néanmoins besoin d'une ou plusieurs traductions d'experts, faites manuellement et de manière anticipée, qui seront considérées lors de l'évaluation comme des traductions de référence de la phrase source. Les métriques d'évaluation automatique sont basées sur le postulat que plus l'hypothèse de traduction faite par la machine est proche de celle faite par l'humain, meilleure elle est. Le paradigme fondamental de l'évaluation automatique de la traduction repose donc sur l'idée de comparer l'hypothèse obtenue par un système de traduction automatique à une traduction de référence produite par un humain.

Principe

Les mesures automatiques ont pour but de déterminer le degré de ressemblance entre une hypothèse de traduction émise par le système et la (les) traduction(s) donnée(s) comme référence. La problématique du domaine consiste alors à déterminer au mieux les critères de comparaison utilisés. Dans la littérature, ces critères reposent sur deux aspects principaux : une composante d'appariement qui permet d'assurer que la traduction automatique comporte les « bons » mots (c'est à dire les mêmes mots que ceux de la traduction de référence) et une composante d'ordre permettant d'évaluer la distance et la place d'un même mot entre l'hypothèse et sa référence. Pour cela, les métriques automatiques utilisent deux concepts principaux : le recouvrement en n-grammes et la distance d'édition. Nous présenterons par la suite, une liste non exhaustive des principales métriques utilisées dans le développement de systèmes de traduction automatique probabilistes. Parmi les dizaines de mesures qui ont été proposées dans la littérature, seules quelques unes sont couramment utilisées et considérées aujourd'hui comme les mesures standard d'évaluation automatique dans la plupart des campagnes d'évaluation [8]. Parmi celles-ci figurent : le score WER, le score BLEU, le score NIST, le score METEOR et le score TER.

Métriques reposant sur la distance d'édition

La distance d'édition (aussi appelé distance de Levenshtein) est une notion ancienne, introduite en 1974 dans [Wagner et Fischer 1974]. Elle reste cependant très utilisée, entre autres dans le domaine du traitement automatique du langage naturel, pour déterminer la similarité entre deux chaînes de caractères. La distance d'édition $D(h, r)$ de deux séquences h et r mesure le nombre minimal d'opérations d'édition nécessaires pour convertir la chaîne de caractère h en r. Il existe plusieurs métriques dérivées de la distance d'édition : selon les opérations d'édition qu'elle considère et selon les entitées sur lesquelles elle est calculée (caractères, mots, etc.).

Le score WER Lorsque la distance d'édition est calculée au niveau des mots, elle est appelée « taux d'erreur en mots » ou *Word Error Rate* (WER) en anglais. Cette mesure empruntée au domaine de la reconnaissance vocale est l'une des premières mesures

8. Il est toutefois important de noter que les campagnes d'évaluation de systèmes de traduction n'utilisent les métriques automatiques qu'à titre indicatif et que l'évaluation manuelle constitue souvent l'essentiel des efforts founis lors de ces campagnes.

proposée pour l'évaluation des sorties de traduction automatique [Vidal 1997, Tillmann et al. 1997a]. Le score WER prend en compte le nombre minimal d'opérations d'édition nécessaire pour transformer l'hypothèse de traduction en sa référence. Les opérations permises sont : l'ajout d'un mot, la suppression d'un mot et le déplacement d'un mot. Elle est calculée selon la formule suivante :

$$WER(h,r) = \frac{\sum_{e \in \{supression, ajout, remplacement\}} C_e \times P_e}{taille_r}$$

où C_e représente le nombre d'opération d'édition de type « e » ($e \in$\{supression, ajout, remplacement\}), P_e le poids de l'opération « e » (traditionnellement, les opérations ont toutes un poids égal à 1) et $taille_r$ le nombre de mots dans la phrase de référence. Moins il y a de modifications à effectuer, meilleur est le score ou autrement dit : plus le taux d'erreur est faible (au minimum 0), plus la traduction est bonne. Le taux maximum n'est pas borné et peut dépasser 1 en cas de très mauvaise reconnaissance, en particulier s'il y a beaucoup d'insertions. Cette métrique n'est pas vraiment adaptée au domaine de la traduction automatique car, contrairement au domaine de la reconnaissance vocale où une seule réponse correcte existe, il existe le plus souvent un très grand nombre de traductions correctes différentes et très dispersées au sens de la distance de Levenshtein. Afin de tenter de répondre plus précisément à la problématique de l'évaluation de traductions automatiques, plusieurs variantes de cette mesure ont été depuis proposées comme par exemple : le taux d'erreur indépendant de la position (PER pour *Position-independent Error Rate*) ou le taux d'erreur à références multiples (multi-reference WER) [Nießen et al. 2000].

Le score TER Cette mesure peut être vue comme paradigme d'évaluation lié à l'utilisation des systèmes de traduction automatique comme aide à la traduction manuelle : la qualité d'une traduction est évaluée en fonction du nombre d'éditions (ajout, suppression, remplacement ou déplacement de mots) faites par un correcteur humain pour obtenir une sortie qu'il juge bonne.

La distance d'édition a été adaptée à la problématique d'évaluation de la traduction à travers le score TER (*Translation Error Rate* ou *Translation Edit Rate*) introduit dans [Snover et al. 2006]. Cette métrique est similaire au *Word Error Rate* mais elle permet, en plus des trois opérations d'édition précédemment définies, une opération supplémentaire : le déplacement de segments ou séquences de mots adjacents (*shift* en anglais). A la différence du score WER, le score TER permet de ne pénaliser que modérément une traduction qui contiendrait une chaîne de mots correcte (qui apparaît dans la référence) mais pas à la bonne place. La mesure a été définie pour représenter le nombre minimum d'opérations d'édition effectué par un annotateur humain pour modifier une hypothèse de traduction en une traduction correcte d'un point de vue grammatical et sémantique. Quand la séquence d'édition est déterminée automatiquement entre l'hypothèse et une traduction référence, le score est appelé TER alors que quand la séquence d'éditions est réalisée par un correcteur humain, le score est appelé HTER (pour *Human Targeted Error Rate*). La procédure HTER est donc une méthodologie visant à créer des références « ciblées » (*Human Targeted*) à partir d'un

résultat de traduction automatique. Un annotateur édite l'hypothèse de traduction donnée par le système pour créer une nouvelle traduction de référence de façon à ce qu'elle soit correcte et aussi proche que possible de l'hypoyhèse de traduction. Le protocole a l'avantage de ne pas nécessiter de locuteurs bilingues car seule la traduction de référence est utilisée pour générer la correction de la traduction (le correcteur n'a pas forcément accés à la phrase source ayant généré la traduction).

La métrique TER possède une autre variante, TERp (pour *Translation Edit Rate Plus*), qui permet de prendre en compte, entre autres, les substitutions de segments (en utilisant des tables de paraphrases), la synonymie et les expressions multi-mots [Snover et al. 2009].

Métriques reposant sur le recouvrement en n-grammes

Les n-grammes sont des sous-séquences de n mots consécutifs, construites à partir d'une séquence donnée. L'hypothèse de traduction et la traduction dite de référence sont comparées au niveau des mots mais également au niveau des bi-grammes, tri-grammes, etc. Les métriques reposant sur le recouvrement en n-grammes estiment la fraction de n-grammes présente à la fois dans la phrase à évaluer et dans la phrase de référence.

Le score BLEU Le score BLEU (pour *BiLinguaL Evaluation Understudy*) a été proposé en 2002 dans [Papineni et al. 2002]. Pour calculer le score BLEU entre une traduction candidate c et une traduction de référence r, on utilise généralement la formule I.7 dans laquelle : N représente la taille maximale des n-grammes considérés (il a été montré que le meilleur compromis entre coût et efficacité consiste à utiliser jusqu'aux 4-grammes), et $n\_grammes_t$ et $n\_grammes_r$ sont respectivement les ensembles de n-grammes des phrases cibles t et des phrases de référence r. Le score BLEU est calculé en considérant les statistiques de co-occurrence de n-grammes pour chacun des segments qui sont ensuite sommées sur tous les segments du texte. Cette moyenne est multipliée par une pénalité de brièveté, destinée à pénaliser les systèmes qui essaieraient d'augmenter artificiellement leurs scores de précision en produisant des phrases délibérément courtes.

$$BLEU(t,r) = BP \times \exp\left(\sum_{n=1}^{N} \omega_n \times \log p_n\right) \tag{I.7}$$

où

$$p_n = \frac{count(n\_grammes_t \cap n\_grammes_r)}{count(n\_grammes_t)}$$

L'ensemble $(n\_grammes_t \cap n\_grammes_r)$ désigne les n-grammes communs à la phrase t et à la phrase r et $count(n\_grammes_t)$ désigne le nombre de n-grammes de la phrase t.

Le coefficient ω_n est un poids qui pondère la précision p_n des n-grammes. On se ramène à une moyenne géométrique de ces précisions avec $\omega_n = \frac{1}{N}$. Le facteur BP est

la pénalité de brièveté qui sert à éviter que le score BLEU ne favorise les traductions candidates courtes pour lesquelles $n\_grammes_t$ est petit, ce qui augmente artificiellement le quotient dans l'exponentielle de la formule de BLEU. Quand la phrase cible est plus petite que la phrase de référence, la score BLEU est pénalisé avec un $BP < 1$ qui le fait diminuer. Dans le cas contraire, le score BLEU est pondéré par un $BP = 1$. La pénalité de brièveté s'estime comme suit :

$$BP = min\left(1, \exp\left(\frac{count(word_t)}{count(word_r)}\right)\right)$$

où $count(word_t)$ et $count(word_r)$ sont respectivement le nombre de mots de la phrase cible et celui de la phrase de référence.

En pratique, on estime le logarithme du score BLEU (que l'on multiplie quelques fois par 100) en considérant jusqu'aux quadri-grammes. Son calcul prend alors la forme suivante :

$$\log\left(BLEU(r,t)\right) = min\left(\frac{count(word_r)}{count(word_t)}, 0\right) + 0,25 \times (\log(p_1) + \log(p_2) + \log(p_3) + \log(p_4))$$

Le score BLEU a été conçu pour l'évaluation à l'échelle de corpus et n'est pas adapté lorsqu'il s'agit d'évaluer des phrases seules. Pour ce faire, une adaptation du score BLEU au niveau des phrases a été introduite dans [Lin et Och 2004]. Elle est définie comme la moyenne arithmétique des précisions modifiées n-gramme (avec $1 \leq n \leq 4$) par la formule :

$$BLEU(t,r) = \frac{1}{4} \times \sum_{n=1}^{4} p_n$$

Elle diffère de la formule destinée à l'évaluation à l'échelle des textes dans le sens où elle ne comporte pas de terme pénalisant la brièveté et est linéarisée.

Le score BLEU varie de 0 à 1 et, étant un score de précision, il est d'autant meilleur qu'il est grand. Parmi les métriques usuelles, le score BLEU occupe une place prépondérante dans l'apprentissage discriminant des systèmes de traduction et la mesure a gagné le statut de mesure automatique de référence au sein de la communauté de la traduction automatique. Cependant, ce score est très controversé, notamment dans [Callison-Burch et al. 2006, Turian et al. 2003, Babych et Hartley 2004] sur le fait qu'il évalue et s'appuie sur la ressemblance du résultat avec la traduction de référence et non la qualité de la traduction en elle-même. Une bonne traduction peut obtenir un score très faible sur le simple fait qu'elle diffère de la traduction de référence précisée et inversement. Le score BLEU nécessite généralement plusieurs traductions de référence pour être fiable mais il existe malheureusement peu de corpus à références multiples et l'évaluation avec BLEU est souvent faite en utilisant une référence unique ce qui aboutit souvent à une corrélation faible avec les jugements humains.

Le score NIST Le score NIST a été proposé en 2002 dans [Doddington 2002] par le *National Institute of Standards and Technology*)[9]. Ce score repose sur un principe

9. http ://www.nist.org

similaire à celui du score BLEU et l'adapte légèrement. La modification la plus notable est que, dans le score NIST, les n-grammes sont pondérés par leur quantité d'information, notion déterminée par leur fréquence : les n-grammes rares contribuent plus au score final que les n-grammes fréquents. Par ailleurs, l'expression de la pénalité de brièveté est légèrement différente de celle de BLEU.

Le score METEOR Le score METEOR présenté par [Banerjee et Lavie 2005] en 2005 introduit la notion d'appariement approximatif : contrairement aux scores BLEU et NIST, un mot de l'hypothèse est jugé correct non seulement s'il apparaît dans la référence, mais également si un de ses synonymes ou un mot morphologiquement apparenté y apparaît. Cette généralisation permet de rendre compte d'une partie de la variabilité des traductions possibles. Lors d'une première étape, un algorithme itératif aligne les mots strictement identiques et tente ensuite, dans une seconde étape, d'aligner les mots restants en utilisant un module de dérivation : lors de cette étape, des termes comme *joli* et *jolie* pourront être appariés. De nombreux modes d'appariement approximatif ont été progressivement introduits et notamment en utilisant des paraphrases et des synonymes issus de ressources propres à la langue traitée. Le score final correspond à la combinaison du rappel et de la précision calculée à partir de l'ensemble des appariements et d'un facteur rendant compte de la fragmentation (différence entre l'ordre des mots de la référence et de l'hypothèse).

3.3 Limitation des évaluations

Le nombre important de mesures automatiques et de protocoles de jugements humains sont représentatifs de la difficulté d'évaluer la qualité d'une traduction. Déterminer la qualité d'une traduction est un problème difficile, notamment car il faut tout d'abord être capable de définir ce qu'est une « bonne » traduction et ensuite pouvoir rendre compte de la variabilité des traductions possible. L'exemple de la Figure I.8 illustre la complexité de la tâche d'évaluation due aux nombreuses manières équivalentes de traduire une même phrase source.

Le principal objectif des mesures automatiques est de prédire le plus précisemment possible le jugement d'un expert humain dans le but de se substituer à lui. Cette problématique donne lieu à des axes de recherches consacrés à la méta-évaluation où la qualité d'une métrique est évaluée en regardant sa corrélation avec des scores attribués par des juges humains.

Néanmoins, au jour d'aujourd'hui, étant donnés un texte source et une traduction candidate, seule une personne bilingue, voire connaissant le contexte et les intentions de l'auteur du texte source, peut véritablement juger de la qualité de la traduction candidate. Une évaluation idéalement menée prendrait en considération la tâche de traduction dans sa situation d'usage. La traduction automatique restant un moyen de parvenir à une fin, son utilité devrait être évaluée en tant que telle, en considérant l'aide apportée pour accomplir une tâche donnée, par exemple : produire des traductions de haute qualité par post-édition, collecter de l'information dans une langue étrangère ou encore assister la communication.

FIGURE I.8 – Exemple de 10 traductions humaines en anglais pour une même phrase source chinois (extrait de [Koehn 2011]).

Pour des raisons évidentes, cette évaluation idéale est envisageable mais extrêmement coûteuse à mettre en œuvre. Loin de considérer le contexte de traduction, les métriques automatiques se limitent à évaluer la capacité d'un système à « imiter » des traductions données comme références (même si certaines cherchent à aller plus loin comme par exemple METEOR ou HTER). Et, en pratique, malgré les nombreuses critiques, les études actuelles se contentent souvent de mesures d'évaluation automatique qui donnent des résultats satisfaisants pour comparer, entres autres, rapidement et sans intervention humaine deux systèmes similaires.

4 Du tout automatique au supervisé par l'humain

Bien que la traduction « toute automatique » ait prouvé son efficacité, les outils restent actuellement intrinsèquement limités et la qualité des traductions produites est généralement jugée insuffisante pour être utilisée telle quelle. La tâche de traduction étant alors jugée trop difficile pour la machine seule, l'expertise humaine est invitée à seconder le système.

Même si en pratique et dans l'usage, l'humain a toujours été intégré dans le processus de traduction, ces dernières années, de nombreux travaux se sont appliqués à développer des méthodologies permettant d'améliorer les systèmes en utilisant l'intervention humaine dans un processus incluant l'humain dans la « boucle » de traduction (*human in-the-loop*).

Dans les nombreuses études qui s'intéressent aux protocoles de traduction où interragissent machines et humains, l'humain qui était jusqu'ici vu comme « utilisateur » devient alors aussi « contributeur ».

Sur l'axe de la collaboration homme-machine, les études présentes dans la littérature

se distinguent maintenant selon deux approches : la traduction automatique au service de l'humain (le système est vu comme un support d'aide à la tâche effectuée par l'humain) et l'humain au service de la traduction automatique (l'humain intervient pour aider l'apprentissage du système).

4.1 La traduction automatique au service de l'humain

La collaboration homme/machine consiste ici à utiliser le système automatique pour répondre à un besoin de l'humain : l'humain effectue la tâche de traduction avec un support informatique pour aide.

Traduction assistée par ordinateur

Dès l'apparition des premiers outils informatiques, les progrès technologiques et les impératifs économiques ont poussé les services de traduction professionnelle à travailler avec les machines.

Parallèlement aux recherches en traduction automatique, est apparu tout naturellement dans les agences de traduction, un nouveau modèle de travail qui concilie l'automatisme et la rapidité de la machine à l'expertise du traducteur professionnel dans le but de mieux répondre aux besoins du marché.

L'ensemble des technologies qui visent à aider les traducteurs professionnels en mettant à leur disposition des outils informatiques destinés à faciliter la tâche de traduction est appelé Traduction Assistée par Ordinateur (TAO). Outre les fonctions permettant la gestion des projets de traduction (analyse statistique, gestion du temps, recherche de termes, etc.), les logiciels existants proposent divers outils pour aider à la traduction. Ceux traditionnellement utilisés sont :

- les mémoires de traduction (couples de phrases pré-traduites de façon antérieure) ;
- les bases de données terminologiques ;
- les systèmes de traduction automatique.

Dans un scénario de TAO où le texte à traduire n'est pas contrôlé par une grammaire limitée et un vocabulaire spécifique à un domaine, la traduction de haute qualité et entièrement automatique est difficilement envisageable. Pour parvenir à une qualité de traduction « publiable », le traducteur a la possibilité d'intervenir à différents niveaux du processus de traduction automatique :

- avant la traduction automatique : il s'agit d'une pré-édition où le texte source est « préparé », normalisé ou désambiguisé en vue de sa traduction automatique ;
- pendant la traduction automatique : on parle de traduction interactive ;
- après la traduction automatique : il s'agit de post-édition où le texte cible produit par le système de traduction est « révisé » ou corrigé.

Pré-édition

Le format et la qualité linguistique d'un texte donné en entrée d'un système de traduction automatique influe directement sur la qualité du résultat. Si un texte source

contient des fautes de syntaxe ou d'ortographe, et/ou que son format de présentation le rend illisible ou incompréhensible, sa traduction automatique sera, très probablement, de mauvaise qualité.

La pré-édition est une action qui consiste à agir en amont du processus de traduction automatique, en préparant un texte en vue de sa traduction par un système automatique. La phrase source est alors modifiée avec pour objectif d'améliorer la qualité du rendu du système.

Cette opération a pour objectif d'adapter le texte source soit dans son contenu, soit dans sa forme et peut impliquer plusieurs types de tâches : modification du format du texte, correction des erreurs linguistiques, orthographiques ou typographiques, simplification des structures syntaxiques complexes ou ambigües, etc.

Ce procédé est généralement utilisé lorsqu'un document est destiné à être traduit en plusieurs langues.

Traduction interactive

Comme nous avons pu le constater, notament à travers les exemples proposés dans ce manuscrit, la production de traductions automatiques de haute qualité est limitée par l'ambiguïté et la complexité inhérente de la langue naturelle.

Un des axes de recherche qui tente de résoudre ces problèmes consiste à solliciter l'utilisateur d'un système de traduction automatique, au cours du processus lui-même, dans le but de « guider » la production du résultat : la traduction est alors dite interactive.

Les systèmes interactifs ont recours à l'expertise humaine soit pour résoudre les ambiguïtés de la langue (on parle de désambiguisation interactive), soit pour co-construire la traduction (on parle alors de traduction interactive par auto-complétion).

La recherche en désambiguisation interactive est apparue avec les premiers systèmes de traduction automatique, dans les années 60 avec les projets MIND [1963-1973] [Kay 1973] et IST [1973-1981] [Melby 1981] puis s'est développée plus tard dans le projet LIDIA [Blanchon et Boitet 2007].

La désambiguisation interactive d'un énoncé peut être d'ordre syntaxique (on cherche à associer une catégorie syntaxique à un mot ou groupe de mots) ou d'ordre sémantique (on cherche à clarifier le sens d'un mot). Dans un premier temps, l'énoncé à traduire est analysé par un module automatique, et, s'il est identifié comme ambigü, l'utilisateur est invité à choisir parmi les différentes interprétations proposées par le système. Ce dialogue homme/machine peut être implémenté sous la forme de questions, de propositions de traductions ou encore de séquences à étiqueter. Les réponses fournies par l'utilisateur permettent au système de poursuivre la traduction de manière autonome et d'aboutir à une traduction de qualité.

D'autres outils de traduction interactive utilisent un modèle de prédiction pour suggérer une nouvelle proposition de traduction au fur et à mesure où l'annotateur modifie la traduction initiale. Le principe utilisé est celui de l'auto-complétion interactive de la traduction : à chaque modification de mot (ou caractère) faite par le traducteur sur l'hypothèse de traduction, le système cherche de nouveau la meilleure hypothèse de traduction pour le reste de la phrase en supposant le segment précédant le mot (ou

caractère) modifié comme correct. A noter qu'il est ici nécessaire pour l'annotateur d'effectuer les corrections du texte de gauche à droite. Les premiers travaux portant sur la traduction interactive par auto-complétion (ou IMT pour *Interactive Machine Translation*) sont issus du projet Transtype [10] financé par le gouvernement Canadien de 1997 à 2000 [Langlais et al. 2000], puis du projet Transtype2 financé par la commission Européenne de 2002 à 2005 [Barrachina et al. 2009]. L'approche a été implémentée par l'Université d'Edimbourg dans l'interface de TAO Caitra [Koehn 2009].

Post-édition

L'association des technologies de traduction assistée par ordinateur et de traduction automatique a, au cours de son évolution, donné lieu à une nouvelle activité et spécialisation : la post-édition.

La post-édition désigne l'activité consistant à réviser un texte traduit automatiquement afin de s'assurer qu'il soit intelligible (en transmettant le sens de la phrase source) et qu'il respecte les règles de grammaire, syntaxe et orthographe de la langue ciblée. La personne chargée d'effectuer cette tâche est appelé un post-éditeur.

Utilisée dans un contexte de traduction professionnelle, la post-édition a pour objectif d'augmenter la productivité (voire la qualité) des traducteurs, en combinant la rapidité d'exécution de la machine avec l'expertise de l'humain.

L'intégration progressive des technologies de traduction automatique dans l'environnement des traducteurs professionnels conduit à penser que le marché de la traduction s'oriente résolument vers la post-édition. Cette constatation est confirmée par les récents choix technologiques faits par des acteurs majeurs de la traduction professionelle (entreprises d'éditeurs de TAO) qui s'accordent à collaborer avec des entreprises spécialisées dans la traduction automatique : en mai 2009, Systran intègre la technologie de TAO Multicorpora à l'un de ses produits de TA (Enterprise Server 7) ; en juin 2009, Google offre un logiciel en ligne permettant de combiner les technologies de TAO et TA (Google Translation Toolkit) ; et en mars 2010, SDL Trados intègre deux nouveaux moteurs de traduction automatique à son logiciel de TAO (SDL Trados Studio 2009 SP). Au même moment (novembre 2010) et dans la même perspective, l'Office Européen des Brevets (EOB) et l'entreprise Google signent un protocole d'accord sur la traduction des brevets. Le système de traduction automatique conçu par Google sera donc dorénavant utilisé pour traduire les brevets délivrés en Europe dans les langues des 38 états membres de l'OEB, et en retour, Google utilisera les brevets traduits et post-édités par l'OEB en vu d'améliorer son système de traduction automatique. Tous les outils cités ont une chose en commun : ils associent TA et traduction professionnelle et impliquent une activité humaine de post-édition.

4.2 L'humain au service de la traduction automatique

A l'opposé de l'approche mise en avant précédemment, nous développons, par la suite, la collaboration homme/machine qui consiste à faire appel à l'expertise humaine pour aider à l'apprentissage d'un système de traduction automatique.

10. http ://www-rali.iro.umontreal.ca/ProjetTransType.en.html

Annotation de données

La plupart des systèmes de traitement automatique du langage naturel utilisent des techniques d'apprentissage supervisées pour créer leurs modèles : cela signifie que les algorithmes d'apprentissage nécessitent des données annotées pour apprendre.

Les annotations peuvent être de plusieurs natures car elles dépendent de la tâche et de son paradigme.

Ces annotations sont utilisées par la machine comme un « oracle », c'est-à-dire une information émise par une personne qui fait autorité et qui n'admet ni le doute, ni la contradiction.

En traitement automatique du langage naturel, le procédé d'annotation nécessite, la plupart du temps, des experts linguistes et consiste à annoter manuellement un ensemble d'exemples non annotés. Une telle campagne d'annotation peut vite se révéler très coûteuse en terme de temps et de ressources mais les données annotées représentent un pré-requis indispensable pour l'apprentissage de la plupart des algorithmes et des annotations linguistiques de qualité sont souvent essentielles afin d'atteindre des performances à l'état de l'art.

L'expertise humaine étant coûteuse et les données manuellement annotées rares, il convient de développer des méthodes adaptées pour utiliser, au mieux, ces précieuses données.

Rétro-action d'utilisateurs

Aujourd'hui, les corpus annotés dédiés aux systèmes automatiques sont relativement peu nombreux et les corpus existants ne couvrent qu'un nombre restreints de paires de langues et/ou domaines. On peut également noter l'usage courant de certains corpus de grande échelle qui proviennent de sources de données dont la collecte est décorrélée de l'usage scientifique (*Europarl*, par exemple, un des corpus largement utilisé pour l'apprentissage de systèmes de traduction automatique probabilistes, est issu de la transcription de débats des assemblées du Parlement Européen).

D'un autre côté, le nombre grandissant d'utilisateurs de systèmes de TA (notamment dû à leur accessibilité en ligne) génère de grandes quantités d'information contenant des retours directs sur les résultats des systèmes. En prenant en compte ces considérations, un nouvel axe de recherche s'est récemment développé autour de l'exploitation des rétro-actions des utilisateurs permettant ainsi de pallier le manque de ressources et tirer bénéfice des données produites par les utilisateurs.

L'apprentissage automatique par rétro-action, désigné couramment par le terme anglais *feedback*, peut être vu comme une collaboration mutuelle et bi-latérale entre l'Homme et la machine. Le mot anglais *feedback* (qui vient de « *to feed* » et « *back* ») peut être traduit littéralement par l'action de « nourir en retour ». L'encyclopédie en ligne Wikipédia définit la rétro-action, au sens large, comme l'action en retour d'un effet sur le dispositif qui lui a donné naissance. Elle se veut être une réponse à l'objectif correcteur.

Dans le domaine de la traduction automatique, l'idée est que les défauts actuels des logiciels peuvent être compensés par les utilisateurs eux-mêmes. Cela nécessite

d'une part de collecter les retours des utilisateurs et d'autre part de ré-intégrer ces informations de façon à enrichir le système.

Collecte de post-éditions de traductions automatiques

Malgré les récents progrès, les systèmes de traduction automatique produisent des traductions qui sont, la plupart du temps, de qualité insuffisante et qui nécessitent une révision plus ou moins importante de la part de l'utilisateur. Dans ce contexte, les outils traductifs sont traditionnellement utilisés pour obtenir rapidement une « pré-traduction » (un premier jet, une ébauche) qui est ensuite éventuellement modifiée par l'utilisateur. Le procédé d'aide à la traduction (qu'elle soit professionnelle ou non) est donc susceptible de produire de gigantesques corpus d'hypothèses de traduction annotées, par exemple, de leurs corrections. Ces données provenant d'expertises humaines constituent une source d'information précieuse pour l'apprentissage, le développement et l'évaluation des systèmes automatiques.

Au delà de l'environnement de la traduction professionnelle, où l'usage de plus en plus démocratisé de la post-édition génère un nombre important de traductions automatiques annotées, la collecte de retours d'utilisateurs commence à être envisagée par le biais des services de traduction omniprésents sur le Web. En effet, devant l'engouement et le succès du Web collaboratif où la participation et la collaboration des internautes est mise à l'honneur (comme en témoigne, par exemple, l'encyclopédie Wikipédia [11]), de nombreux services de traduction en ligne n'hésitent pas à reccueillir les contributions des utilisateurs. En effet, il est de plus en plus souvent possible, sur les sites de traduction en ligne, d'indiquer si l'on est satisfait d'une traduction donnée et/ou de la corriger. De telle campagnes de collecte ont été (ou sont actuellement) effectuées sur les interfaces de Google Translate [12], Bing Translator [13], et Reverso [14]. Bien que disponibles en très grand nombre, les jugements ainsi collectés ne sont pas forcément de bonne qualité puisqu'ils n'ont pas nécessairement été émis par des experts et qu'il est possible que certains utilisateurs aient, volontairement ou non, porté des jugements faux. Savoir si leur nombre peut compenser leur qualité, parfois soupçonnée douteuse, constitue aujourd'hui un axe de recherche à part entière (voir par exemple [Snow et al. 2008]).

Exploitation de post-éditions

La collecte de post-éditions en tant que retours d'utilisateurs et corrections d'hypothèses de traductions commence à intéresser la communauté du domaine de la traduction automatique probabiliste. Elle se fait, en partie en ligne, par le biais d'internautes volontaires. Cependant, il existe à ce jour peu de travaux portant sur l'utilisation de post-éditions pour corriger et améliorer un système de traduction automatique et,

11. http ://fr.wikipedia.org
12. http ://translate.google.fr
13. http ://www.bing.com/translator
14. http ://www.reverso.net/text_translation

à notre connaissance, aucun système capable de s'adapter directement par *feedbacks* d'utilisateurs.

Nous pouvons noter ceux qui utilisent des post-éditions manuelles de systèmes de traduction automatique afin de créer un post-éditeur automatique qui permet de « corriger » les sorties du système [Simard et al. 2007a, Simard et al. 2007b]. De son côté, le projet FAUST (Feedback Analysis for User adaptive Statistical Translation) organise la collecte et l'analyse de *rétro-actions* d'utilisateurs du site de traduction en ligne Reverso dans le but de développer des techniques pour les exploiter [Déchelotte 2010].

Le travail présenté dans ce manuscrit a pour objectif, à long terme, de développer des techniques d'apprentissage rétro-actif pour améliorer de façon interactive et itérative un système de traduction automatique probabiliste à l'aide d'expertises humaines collectées sous forme de post-éditions.

Chapitre II

Création d'un système de traduction probabiliste de référence

1 Choix du contexte applicatif

Le choix du cadre applicatif a été motivé par la volonté de créer un système générique ayant pour objectif principal de valider la méthode proposée. Le système effectuera des traductions de la langue française vers la langue anglaise. Ces langues ont été sélectionnées pour la facilité à trouver des locuteurs natifs ou ayant une pratique avancée de la langue et, d'autre part, pour la place prépondérante qu'elles occupent sur le Web ce qui a pour avantage de simplifier la constitution des corpus représentant ce couple de langues. De la même façon, le domaine d'application choisi est celui des brèves journalistiques, aussi appelées « news » ou dépêches. Ce domaine est défini comme « général » par la communauté du traitement automatique du langage naturel.

2 Corpus

Une des étapes importantes dans la création d'un système de traduction automatique statistique est la collecte des corpus nécessaires à l'apprentissage et au test du système. L'approche empirique nécessite, en effet, l'utilisation de corpus monolingues et de corpus bilingues alignés de taille et de qualité suffisante pour le domaine et la paire de langue envisagés.

2.1 Caractéristiques des corpus

Etant donné notre cadre applicatif, les corpus doivent nécessairement comporter des données journalistiques de type « news » afin de respecter une adéquation suffisante avec l'application visée : style, période, thèmes abordés, etc. Pour cette raison, nous nous devons d'attacher une attention particulière à la période couverte par les corpus et de choisir des données relativement contemporaines pour pouvoir couvrir l'actualité récente. D'autre part, il semble nécessaire de respecter la chronologie des données et

veiller à ce que le corpus d'apprentissage soit antérieur au corpus de test. Enfin, les corpus doivent atteindre une taille suffisante pour permettre des traitements statistiques fiables et une bonne estimation des probabilités.

2.2 Source des corpus

Nous avons donc choisi d'utiliser des corpus issus de la campagne d'évaluation WMT (Workshop on statistical Machine Translation). Ce choix a été guidé par la volonté de pouvoir situer les résultats de notre étude par rapport à d'autres travaux à l'état de l'art.

WMT [15] est une campagne d'évaluation organisée chaque année depuis 2006 avec le soutien de la commission européenne et le projet européen EuroMatrixPlus [16]. A l'origine (2006 et 2007), la campagne qui se focalisait sur la traduction automatique entre cinq langues européennes (le français, l'allemand, l'anglais, l'espagnol et le tchèque) s'est peu à peu diversifiée et a intégré une tâche d'évaluation de résultats de traductions automatique (2008) puis une tâche de combinaison de systèmes (2009) pour enfin proposer une tâche d'estimation de confiance (2012). Soucieuse d'être au plus proche des défis scientifiques contemporains, la campagne n'hésite pas à confronter les participants aux problématiques d'actualité. Ainsi, en 2011, en réponse au tremblement de terre de janvier 2010 à Haïti, la campagne a proposé une tâche de traduction de SMS du créole haïtien vers l'anglais. L'évaluation des participants à WMT se fait sur la base de plusieurs métriques automatiques et de jugements humains (les modalités de l'évaluation sont variables selon les années). Les jugements sont effectués par les participants eux-mêmes (chaque site participant s'engage à faire 8h d'évaluation manuelle). Les organisateurs ont également proposé, en 2010, de collecter à très faible coût des jugements dont la qualité n'est pas contrôlée, en utilisant la plateforme Amazon Mechanical Turk [17].

Les corpus mis à disposition dans le cadre de la campagne d'évaluation WMT sont considérés comme des corpus de référence dans le domaine de la traduction automatique probabiliste.

2.3 Description des corpus

Les corpus sélectionnés pour l'élaboration de notre système de référence sont décrits dans le tableau II.1.

L'apprentissage du système de traduction se base sur deux corpus bilingues alignés de natures différentes. Le premier (ligne (1) dans le tableau II.1) représente des transcriptions de débats issus des assemblées du Parlement Européen (discours portant sur le fonctionnement interne de l'Union Européenne) [Koehn 2005a]. Le deuxième (ligne (2) dans le tableau II.1) contient des données journalistiques qui représentent des éditoriaux, de sujets de politique, économie ou sciences, issus de divers sites Web journalistiques dont le contenu est publié en plusieurs langues. Les sources utilisées sont, entre

15. http ://www.statmt.org/wmt11/
16. http ://www.euromatrixplus.net/
17. https ://www.mturk.com/mturk/welcome

autres, les sites Web : *Project Syndicate*[18], *Libération*[19] et *Le Figaro*[20]. L'ensemble du corpus d'apprentissage du système comprend en tout 1 640 463 énnoncés.

Le modèle de langage du système est appris sur un corpus monoligue Anglais (langue cible) de 48 653 884 phrases. Ce corpus (ligne (3) dans le tableau II.1) comprend les données en langue anglaises des corpus d'apprentissage bilingues ainsi que des documents officiels de l'assemblée générale des Nations Unies[21] [Rafalovitch et Dale 2009].

Les corpus utilisé pour le développement et le test du système (ligne (4) et (5) dans le tableau II.1) sont constitués d'articles journalistiques de la même origine que le corpus d'apprentissage journalistique (ligne (2) dans le tableau II.1) mais couvrent des périodes différentes : de novembre à décembre 2007 pour le corpus de développement et de fin septembre à mi-octobre 2008 pour le corpus de test.

Ces différents ensembles de données bilingues sont le fruit du travail de traducteurs professionnels ou de locuteurs bilingues. Les traductions qu'ils contiennent sont vérifiées de façon interne afin de satisfaire les exigences de qualité de l'organisme où elles sont produites.

Utilisation	Nature du corpus			Taille
	Langue(s)	Origine	Période	(en phrases)
(1) Apprentissage	Anglais/Français	Parlement européen	2007	1 585 819
(2) Apprentissage	Anglais/Français	Sites Web	2006-2007	54 644
(3) Apprentissage	Anglais	Nation Unies	2007	47 013 481
(4) Développement	Anglais/Français	Sites Web	2007	1 200
(5) Test	Anglais/Français	Sites Web	2008	2 852

TABLE II.1 – Description des corpus utilisés pour le système de traduction de référence

3 Apprentisssage du système de référence

Le système de référence créé est un système de traduction statistique à base de segments (ou *phrase-based* en anglais) où les unités de traduction sont les segments (suite de n mots consécutifs). Il repose sur une modélisation log-linéaire pondérée de 14 fonctions de caractéristiques apprise à l'aide de plusieurs techniques et outils disponibles sous licence GPL[22] dans la boîte à outils Moses.

3.1 La boîte à outils Moses

La boîte à outils Moses[23] [Koehn et al. 2007] regroupe des outils et des scripts qui implémentent des techniques permettant de créer et tester des systèmes de tra-

18. http ://www.project-syndicate.org
19. www.liberation.fr
20. www.lefigaro.fr
21. www.uncorpora.org
22. GPL = Gnu Public Licence
23. http ://mosesdecoder.sourceforge.net/download.php

duction probabilistes reposant sur un modèle log-linéaire. C'est l'un des outils *open source* les plus utilisés dans les institutions académiques comme support de base pour l'apprentissage, le développement et le test de systèmes de traduction automatique à base de segments dédiés à la recherche. L'outil Moses a été conçu en 2006, à l'Université d'Edinbourg, par une équipe dirigée par Philipp Koehn et a été principalement programmé par Hieu Hoang. Il est régulièrement mis à jour par des corrections de dysfonctionnement ou l'ajout de nouvelles fonctionnalités.

3.2 Normalisation des corpus

Au préalable à toute utilisation, les corpus utilisés ont été traités afin d'uniformiser la forme des données textuelles qu'ils contiennent et de les rendre plus facilement exploitables par les scripts d'apprentissage proposés dans Moses.

Afin de réduire considérablement la taille du vocabulaire du système, la première opération de pré-traitement de nos données consiste à décapitaliser les textes source et cible. Dans un second temps, nous avons segmenté les textes en unités de référence (étape dite de "*tokenisation*"). L'objectif est ici de transformer les chaînes de caractères en des séquences de "tokens" ou symboles. Cette opération s'applique aux textes sources et cibles et prend en compte les espaces pour séparer les mots, les nombres et les signes de ponctuation.

Au vu des résultats de premières expérimentations, nous avons choisi de traiter en plus, une forme particulière issue de la syntaxe Française : le *t* euphonique. Le *t* euphonique consiste à ajouter la lettre "t" entre un verbe (se finissant par "a", "e" ou "c") et un pronom personnel et à les inverser dans le but d'en faciliter la prononciation. Une séquence est ainsi représentée par la chaine "*verbe*-t-*pronom*" comme par exemple *annonça-t-elle*, *arrive-t-il* ou encore *vainc-t-on*. Cette forme apparaît dans 1,75 % des phrases de notre corpus de test alors que sa fréquence d'apparition est de l'ordre de 0,66 % dans le corpus d'apprentissage. La forme normalisée proposée comporte le pronom sujet en première position suivi du verbe : "*verbe*-t-*pronom*" devient alors "*pronom verbe*". Un exemple de normalisation du *t* euphonique est donné dans la figure II.1. Nos résultats montrent que la modification de cette forme complexe permet un meilleur traitement du texte source français par le système et donc un meilleur résultat de traduction automatique.

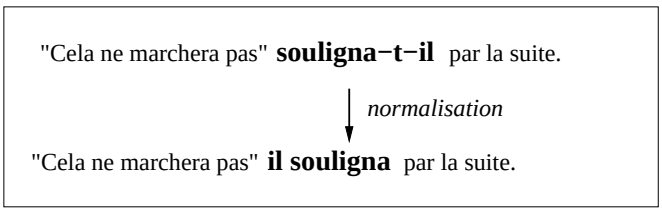


FIGURE II.1 – Normalisation des figures de "t" euphonique

3.3 Modélisation du système de référence

Le système de référence repose sur un modèle log-linéaire composé de 14 fonctions caractéristiques : un modèle de langage de la langue cible ($P(t)$), deux statistiques de traduction sur les segments ($T(t|s)$ et $T(s|t)$), deux probabilités de traduction lexicale sur les mots ($P_{lex}(s|t)$ et $P_{lex}(t|s)$), un modèle de pénalité sur les mots (wp), un modèle de pénalité sur les segments (pp), un modèle de pénalité sur le ré-ordonnencement (d) et six modèles lexicaux de distorsion (D_{m_p}, D_{m_f}, D_{s_p}, D_{s_f}, D_{d_p} et D_{d_f}). Ces fonctions de caractéristiques sont apprises à partir des corpus d'apprentissage monolingues et bilingues (s et t désignent respectivement une phrase source et une phrase cible).

Le modèle de langage de la langue cible

Le modèle de langage de la langue cible (Anglais) de notre système est appris sur notre corpus d'apprentissage Anglais à l'aide de l'outil *SRI Language Modeling Toolkit* (ou SriLM) [Stolcke 2002] qui intègre les techniques de modélisation du langage les plus utilisées par la communauté de la traduction automatique empirique. SriLM permet de construire des modèles n-grammes utilisant différentes techniques de lissage de probabilités.

Pour notre application, nous avons construit un modèle quadri-grammes lissé avec la technique d'interpolation de Kneser-Ney modifiée [Chen et Goodman 1999].

Le modèle de traduction

Le corpus d'apprentisage bilingue aligné phrase à phrase est utilisé pour réaliser des alignements, en mots d'adord avec le logiciel GIZA++ [Och et Ney 2003], puis en segments ensuite avec le logiciel Moses. Les paires de segments textes en langue source et cible du corpus ainsi alignés sont extraites du corpus d'aprentissage afin de créer une table de bi-segments alignés. Cette table de traduction (aussi appelée *phrase-table*) se trouve au cœur du système et est utilisée pour calculer les statistiques des fonctions de caractéristiques bilingues du modèle.

Il est important de noter que la qualité de la traduction produite dépend en grande partie de la quantité et de la qualité des données d'apprentissage de la table de traduction. En effet, le système génère les traductions à partir de cette table qui représente le dictionnaire du modèle. Il est donc nécessaire de posséder un corpus bilingue aligné de qualité pour apprendre le meilleur dictionnaire possible avec de bonnes estimations.

Probabilités de traduction lexicale sur les mots Les poids lexicaux représentés par $P_{lex}(\bar{t}|\bar{s})$ et $P_{lex}(\bar{s}|\bar{t})$ (où $\bar{s}$ et $\bar{t}$ représentent des segments), sont calculés à partir des probabilités issues du corpus d'apprentissage bilingue aligné en mots. Le poids lexical des segments ($\bar{t}$,$\bar{s}$) s'exprime comme suit :

$$P_{lex}(\bar{s}|\bar{t}) = \sum_{n=1}^{N} \left(\prod_{i=1}^{m} \frac{1}{|(i,j)|_{\forall (i,j) \in a_m}} \times \sum_{i} p(s_i, t_j) \right)$$

où N est le nombre d'alignements en mots possibles entre les segments $\bar{s}$ et $\bar{t}$ et :

$$p(s_i, t_j) = \frac{count(s_i, t_j)}{\sum_{s'} count(s'_i, t_j)}$$

où $count(s_i, t_j)$ est le nombre de fois où le mot source s_i a été aligné avec le mot cible t_j dans le corpus d'apprentissage Par exemple, l'alignement de la table II.2 nous donnera la probabilité lexicale suivante :
$P_{lex}(\bar{s}|\bar{t}) = p(s_1, t_1) \times \frac{1}{2}\left(p(s_2, t_2) + p(s_2, t_3)\right) \times p(s_3, null)$

$\bar{s} \setminus \bar{t}$	t_1	t_2	t_3	null
s_1	X			
s_2		X	X	
s_3				X

TABLE II.2 – Exemple d'alignement entre s et t

Statistiques de traduction des segments La statistique de traduction $T(\bar{t}|\bar{s})$ (respectivement $T(\bar{s}|\bar{t})$) représente la probabilité de traduction du segment $\bar{s}$ (respectivement $\bar{t}$) par le segment $\bar{t}$ (respectivemnt $\bar{s}$). Ces statistiques sont issues de la table de traduction entraînée à partir du corpus d'apprentissage bilingue. Leurs formules sont données par :

$$T(\bar{t}|\bar{s}) = \frac{count(\bar{s}, \bar{t})}{\sum_{\bar{s'}} count(\bar{s'}, \bar{t})} \quad et \quad T(\bar{s}|\bar{t}) = \frac{count(\bar{t}, \bar{s})}{\sum_{\bar{t'}} count(\bar{t'}, \bar{s})}$$

où $count(\bar{s}, \bar{t})$ est le nombre de fois où le groupe de mots $\bar{s}$ a été aligné avec $\bar{t}$ dans le corpus d'apprentissage. Ces statistiques permettent d'assurer que le segment en langue source $\bar{s}$ et le segment en langue cible $\bar{t}$ soient une bonne traduction l'un de l'autre.

Le modèle de pénalité sur les mots Cette fonction caractéristique assigne une constante de coût à chaque mot de la phrase cible. Elle permet de compenser la tendance du système à privilégier les phrases courtes en les pénalisant. On s'assure ainsi que les traductions produites ne sont ni trop courtes ni trop longues. Par la suite, on désignera par *wp* le modèle de pénalité sur les mots. Sa valeur est un facteur constant pour chaque mot généré qui par défaut vaut $wp = -1$.

Le modèle de pénalité sur les segments Ce score assigne un poids constant à chaque segment de la phrase cible produite. En faisant varier sa valeur, on peut influer sur la propension du système à utiliser de nombreuses paires de segments (donc courts) ou au contraire lui faire préférer des groupes de mots plus longs. Sa valeur est un facteur constant pour chaque groupe de mots généré qui par défaut vaut $pp = \exp(1) = 2,718$. Par la suite, on désignera par *pp* le modèle de pénalité sur les segments.

La pénalité de ré-ordonnancement La pénalité de ré-ordonnancement évalue le coût linéaire du ré-ordonnancement de la phrase cible. Plus il y a ré-ordonnancement, plus la traduction se révèle coûteuse. Déplacer un segment cible de n segments implique un coût de ω^n (où ω est une constante fixée). Par la suite, on désignera par d la distance de ré-ordonnancement.

Il est à noter qu'il existe un nombre maximal de mouvements qui ne permet pas de déplacement de segment au delà de cette limite. Par défaut, la valeur est fixée à 6 segments et elle ne fait pas partie des paramètres du modèle réglés lors de l'apprentissage.

Les modèles lexicaux de distorsion Les modèles lexicaux de distorsion estiment le ré-ordonnancement des segments, c'est-à-dire le déplacement de segments pour chaque paire de phrase (s, t), proportionnellement à la quantité de segments non déplacés (dits monotones). Ils sont utilisés pour prendre en compte le fait que certains segments sont ré-ordonnancés plus fréquement que d'autres. L'adjectif *extérieur*, par exemple, est mis devant le nom lorsqu'il est traduit en anglais. Les modèles de distorsion considèrent trois types d'orientations : monotonie (qui ne présente pas de ré-ordonnancement), permutation *(Swap)* et discontinuité désignées respectivement par m, s et d. Pour chaque type d'orientation, on considère les voisins du segment traité, c'est-à-dire, le segment actuel a_i en fonction du précédent a_{i-1} et en fonction du suivant a_{i+1}. Etant données une phrase source s, une phrase cible $t = (\bar{t}_1, \bar{t}_2 \ ... \ \bar{t}_M)$, toutes deux décomposées en groupes de mots, et un alignement $a = (a_1, a_2 \ ... \ a_M)$ qui définit un groupe de mots source $\bar{s}_{a_i}$ pour chaque groupe de mots traduit $\bar{t}_i$, ces modèles estiment la distorsion, où chaque $o_m \in \{monotone, swap, discontinue\}$, comme suit :

$$p(o_p|s,t) = \prod_{i=1}^{M} P(o_i|\bar{t}_i, \bar{s}_{a_i}, a_{i-1}, a_i) \quad et \quad p(o_f|s,t) = \prod_{i=1}^{M} P(o_i|\bar{t}_i, \bar{s}_{a_i}, a_{i+1}, a_i)$$

Ces probabilités sont apprises lorsque l'on extrait les paires de segments à partir de l'alignement en mots :

$$p(o|s,t) = \frac{count(o,s,t)}{\sum_{o \in (m,s,d)} count(o,s,t)}$$

Ces 6 fonctions de caractéristiques (chacune correspondant à une orientation qui prend en compte le segment précédent p ou le segment suivant f), seront désignées par la suite par : D_{m_p}, D_{m_f}, D_{s_p}, D_{s_f}, D_{d_p}, D_{d_f}.

Optimisation des poids du modèle

Par défaut l'outil Moses affecte des valeurs prédéfinies aux poids des 14 fonctions caractéristiques du modèle log-linéaire. Ces valeurs par défault sont génériques, c'est pourquoi il est souvent recommandé d'ajuster le modèle afin d'obtenir des poids plus adaptés à la paire de langue et au domaine des corpus traités. Il est à noter que cela n'est pas toujours vérifié en pratique, il arrive que l'algorithme d'ajustement des poids

ne parvienne pas à améliorer significativement le score obtenu avec les valeurs par défaut.

Dans nos expérimentations, cet ajustement des poids du modèle est fait avec l'algorithme *Minimum Error Rate Training* [Och 2003] implémenté dans la boîte à outils Moses.

Décodage

Lors du décodage, le système créé prend en entrée un corpus en langue source (ici francais) et en fournit une traduction phrase à phrase en langue cible (ici anglais). Le système utilisant une approche log-linéaire, l'algorithme de décodage cherche la phrase cible la plus probable étant donnée une phrase source, par maximisation de la somme des scores des 14 fonctions de caractéristiques $h_m(t, s)$ présentées précédemment :

$$t^* = argmax_t p(t|s) = argmax_t \sum_{m=1}^{14} \lambda_m h_m(t, s)$$

Cette opération de décodage fait partie des implémentations de l'outil Moses et l'algorithme utilisé est celui de la recherche en faisceau (ou *beam search* en anglais).

4 Validation du système de référence

4.1 Participation à la campagne d'évaluation WMT 2010

Il existe de nombreuses campagnes d'évaluation en traduction automatique. En plus de favoriser les échanges scientifiques et le travail coopératif, elles permettent d'évaluer la qualité de différents systèmes et de confronter les approches des différents participants.

Notre système de traduction de référence a été soumis au jugement de la communauté scientifique lors de la participation du Laboratoire d'Informatique de Grenoble à la campagne d'évaluation WMT 2010 [Potet et al. 2010]. Les performances obtenues lors de cette campagne ont montré que le système pouvait être considéré comme étant à l'état de l'art.

4.2 Evaluation du système

Par la suite, le score du système de référence sera exprimé en terme de score BLEU (ramené à un pourcentage) calculé à l'aide de l'implémentation de la métrique automatique proposée par l'outil MTEval version 13. Il est à noter que, ne disposant pas de plusieurs références pour une même phrase source du corpus de test, le score BLEU sera calculé à partir d'une référence unique.

Les performances de notre système de référence sur les corpus de développement et de test (avec les poids par défaut et avec les poids ajustés à l'aide de la technique MERT) sont présentées dans le tableau II.3. La performance du système sur le corpus de test (Score BLEU de 25,27) servira de score de référence lors des expérimentations

liées à son optimisation et les traductions qu'il produit seront nommées par la suite *hypothèses de traduction*.

Ces résultats ont permis de classer ce système de référence parmi les 10 meilleurs participants à la campagne d'évaluation internationale WMT 2010 pour la tâche de traduction du français vers l'anglais[24] [Callison-Burch et al. 2010].

Corpus	score BLEU
Developpement	24.32 (24.50)
Test	**25,27** (25.05)

TABLE II.3 – Performance du système de référence avec les poids par défaut (avec les poids ajustés à l'aide de la technique MERT)

4.3 Significativité des différences entre deux scores BLEU

La significativité des différences entre les scores BLEU des expérimentations présentées dans ce manuscrit est évaluée selon la technique d'amorce par ré-échantillonage (ou *bootstrap resampling method* en anglais) proposée dans [Koehn 2004]. La méthode étant coûteuse à mettre en œuvre pour chacun des résultats obtenus, nous utilisons les résultats expérimentaux de l'étude de P. Koehn pour estimer le taux de confiance à accorder aux variations de scores BLEU. Le taux de confiance à accorder à une différence (en pourcentage) entre deux scores BLEU, en tenant compte de la taille du corpus de test, est donnée par le tableau 1.1 de l'annexe 1 page 134.

D'après ce tableau, en tenant compte de la taille de notre corpus de test (environ 3 000 phrases) et du score BLEU obtenu par le système de référence (25,27), il est nécessaire d'obtenir une différence absolue de 0,38 points BLEU (ou relative de 1,5 %) avec un second score pour que celle-ci soit jugée statistiquement significative à 95 %, avec une certitude de 100 %.

Des exemples de traductions produites par le système de référence sont donnés dans le tableau II.4. Nous constatons que, pour les exemples 2., 3. et 4., l'hypothèse de traduction produite par le système est, contrairement à la traduction de référence, une très bonne traduction littérale de la phrase source.

24. Le classement des systèmes de traduction a été réalisé sur la base d'une évaluation faite par des traducteurs professionnels. Le rang de notre système dans le classement peut être calculé en prenant en considération le respect des conditions "contraintes" de la tâche c'est à dire l'utilisation exclusive des ressources fournies lors de la campagne (notre système est classé $5^{ième}$ sur 9), ou pas (notre système est alors classé $9^{ième}$ sur 16).

Phrase source	Phrase de référence	Hypothèse de traduction
1. Cette décision cruciale pour l'avenir d'une région instable mettra à l'épreuve la détermination et l'unité occidentales.	This decision crucial to the future of an unstable region will test western determination and unity.	This decision crucial for the future of an unstable region will test the resolve and unity in the west.
2. Le Kosovo indépendant doit pouvoir vivre en sécurité et ses minorités doivent être protégées.	An independent Kosovo must be secured and its minorities protected.	The independent Kosovo must live in security and its minorities must be protected.
3. La BCE souhaiterait maintenir le taux d'inflation au-dessous mais proche de deux pour cent.	The ECB wants to hold inflation to under two percent or somewhere in that vicinity.	The ECB would like to maintain the inflation rate below but close to two percent.
4. Le prix de l'huile alimentaire et des produits laitiers a également considérablement augmenté en deux mille sept.	Consumers also have had to pay significantly more for vegetable oils and dairy products in two thousand seven.	The price of oil food and dairy products has also dramatically raised in two thousand seven.

TABLE II.4 – Exemples de traductions faites par le système de référence

Chapitre III

Expérimentations préliminaires

Nous rappelons que le scénario envisagé est un système de traduction automatique probabiliste qui propose à l'utilisateur de corriger la traduction proposée par le système et qui prend en compte cette correction en la ré-intégrant dans le système de façon à l'améliorer.

1 Objectif de l'étude

Dans un premier temps, nous avons créé un système de traduction automatique probabiliste de référence et validé ses performances comme étant à l'état de l'art. Nous souhaitons maintenant étudier la faisabilité de l'intégration de rétro-actions d'utilisateurs pour améliorer ce système. L'objectif des travaux qui vont suivre est donc d'expérimenter différentes méthodes pour prendre en compte des corrections effectuées manuellement de façon à améliorer le système de traduction.

Pour cela, nous avons d'abord fait appel à des annotateurs humains pour corriger un petit ensemble d'hypothèses de traduction produites par notre système et nous proposons et évaluons des protocoles expérimentaux dont l'objectif est d'intégrer les post-editions collectées au système initial.

2 Post-édition de 175 hypothèses de traduction

Dans notre scénario, la rétro-action de l'utilisateur est une post-édition, c'est-à-dire une hypothèse de traduction donnée par notre système qui a été vérifiée et éventuellement corrigée par un annotateur humain. Nous avons donc collecté des corrections de traductions issues de notre système en faisant post-éditer, par des annotateurs, un petit ensemble de 175 hypothèses de traduction issues de notre système.

2.1 Corpus de post-éditions

La post-édition est une tâche relativement coûteuse en terme de temps et d'efforts d'annotation. Cette étude n'ayant qu'une visée préliminaire, nous avons choisi d'annoter dans un premier temps, un corpus restreint de 175 énnoncés.

Le corpus de post-édition choisi est disjoint de ceux utilisés pour l'apprentissage, le test et le développement du système de traduction de référence. Il contient 175 énnoncés parallèles en anglais et en français. Les textes proviennent de différents sites Web journalistiques (Libération [25], Le Figaro [26], Les Echos [27], etc.) qui se trouvent être déjà traduits par des traducteurs professionnels. Ce corpus est nommé par la suite « PE ».

2.2 Collecte des post-éditions

Nous avons fait traduire le corpus PE par notre système de traduction de référence (présenté dans le chapitre II) puis nous avons fait corriger les hypothèses de traduction du système par des annotateurs volontaires, via l'interface d'annotation $SECTra\_W$ [Huynh et al. 2008]. La consigne donnée est d'effectuer les corrections minimales de façon à ce que l'hypothèse de traduction devienne une traduction correcte de l'énoncé source. Dans certains cas, aucune correction n'est nécessaire. Les correcteurs humains auxquels nous avons fait appel ne sont ni des traducteurs professionnels, ni des locuteurs natifs de la langue anglaise. Ce dernier point, qui peut donner lieu à controverse, sera discuté plus précisément au chapitre suivant. L'intérêt est ici d'obtenir une traduction comportant les corrections nécessaires et suffisantes pour permettre, à un non-natif de la langue source, de comprendre le sens de l'énoncé traduit. Les post-éditions obtenues sont des traductions correctes de l'énoncé source français, aussi proches que possible de celles produites par le système.

2.3 Exemples de post-éditions

Des exemples d'hypothèses de traduction corrigées (ou post-éditions) sont données dans le tableau III.1 Comme nous pouvons le constater, les traductions professionnelles fournies dans le corpus parallèle (aussi appelées *gold standard*) se révèlent être des traductions très libres, parfois éloignées de l'énoncé source (contenant même quelques fois des erreurs) alors que les hypothèses de traduction corrigées sont proches des sorties du système et semblent être propices à être utilisées en vue de le corriger. Par la suite, lorsque cela est nécessaire, nous préciserons si l'on utilise le corpus parallèle contenant comme référence les traductions professionnelles fournies avec le corpus d'origine « PE_{std} » ou les hypothèses de traduction de notre système de référence corrigées manuellement par nos soins « PE_{corr} ».

25. www.liberation.fr
26. www.lefigaro.fr
27. www.lesechos.fr

Enoncé source	*Gold standard*	Hypothèse de traduction	Post-édition
• Les pierres sont sales.	• The ston**e is** dirty.	• The stones are **vile**.	• The stone**s are** dirty.
• J'ai lu cela dans mes mangas.	• I read **about that** in my mang**a**.	• **I have read it in my mangas**.	• **I have read it in my mangas**.
• Il eu de la peine à obtenir un oscar.	• **And awarding him with an oscar had been quite hard**.	• It was hard to get an oscar.	• It was hard **for him** to get an oscar.

TABLE III.1 – Exemples extraits de l'ensemble de 175 post-éditions.

3 Intégration des post-éditions au système de traduction

La collecte précédente nous fournit un corpus de 175 hypothèses de traductions issues de notre système et corrigées par des annotateurs volontaires. Nous proposons et évaluons, par la suite, des techniques pour exploiter au mieux ces données corrigées, en les intégrant dans notre système dans le but de l'améliorer. La figure III.1 présente un schéma de l'intégration des énoncés corrigés dans le système de traduction. Ces données sont injectées à trois niveaux différents du processus de traduction :

- pour enrichir le corpus d'apprentissage : nous proposons d'ajouter les traductions post-éditées au corpus d'apprentissage pour ré-apprendre un nouveau système de traduction ;
- pour ajuster les poids du modèle de traduction de référence : nous proposons d'utiliser les traductions post-éditées pour ajuster les poids du modèle log-linéaire de notre système de traduction ;
- pour corriger automatiquement les sorties du système : le corpus de traductions corrigées est ici utilisé pour modifier les sorties du système par le biais d'un post-éditeur statistique.

Nous évaluerons l'apport des méthodes sur le corpus de post-éditions « PE » (175 énoncés) et sur le corpus de test « $TEST$ » du système de traduction de référence présenté dans le chapitre II (contenant 2 852 énoncés).

3.1 Ajout des post-éditions au corpus d'apprentissage du système

L'idée est d'ajouter le corpus de traductions corrigées au corpus d'entraînement du système de traduction de référence.

Protocole expérimental

Le problème est que la quantité d'énoncés corrigés (175 énnoncés) est infime au regard du corpus initial d'apprentissage (1 640 463 énnoncés). Pour donner plus de poids à ces données corrigées, nous les avons dupliquées 10, 100 et 1000 fois avant de les intégrer au corpus d'apprentissage. Nous créons donc un corpus avec les données

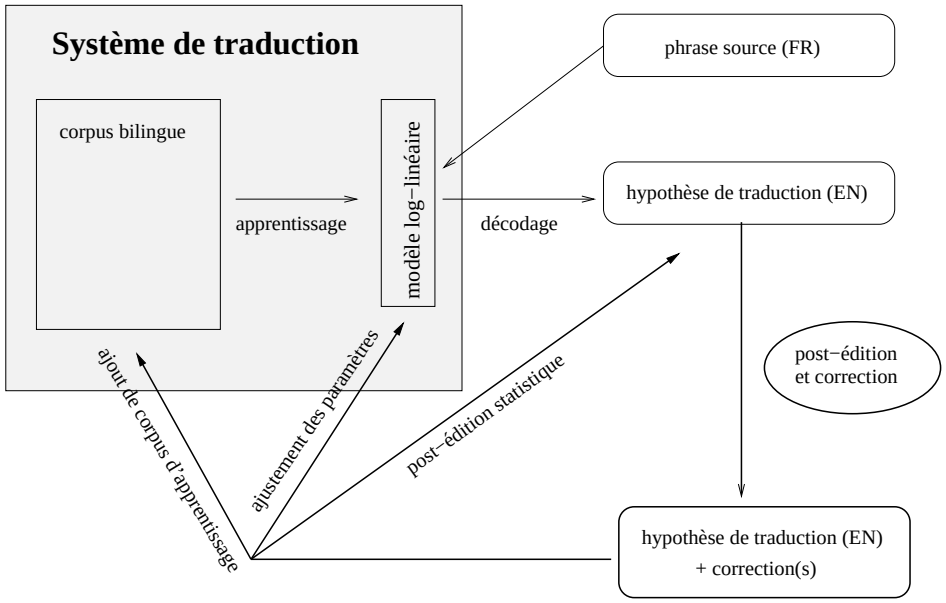


FIGURE III.1 – Schéma de l'intégration des énoncés corrigés dans le système de traduction de référence

d'apprentissage du système de référence auquel on ajoute n fois les données corrigées ($n \in \{1, 10, 100, 1000\}$) : n peut être interprété comme le poids imparti aux données corrigées.

Résultats

La qualité des résultats des systèmes appris est donnée en terme de scores BLEU dans le tableau III.2 sur les corpus PE et $TEST$. La colonne *« Enoncés ≠ »* présente la proportion d'énoncés dont la traduction produite par le système *Baseline* + n * PE_{corr} est différente de celle produite par le système de référence (correspondant à la première ligne, $n = 0$).

Si l'on ajoute le corpus corrigé de 175 énnoncés, en le dupliquant 1000 fois (ce qui revient à augmenter le corpus d'apprentissage de 11 %), 90 % des hypothèses de traductions du corpus PE sont modifiées et l'on passe d'un score BLEU de 23,50 (avec $n = 0$) à 25,73 (avec $n = 1\ 000$). Ceci représente un gain significatif de 10 % (selon [Koehn 2004]). Le gain est plus modeste mais également observable sur le corpus $TEST$, qui passe d'un score BLEU de 25,27 ($n = 0$) à 25,51 ($n = 1000$) avec 65 % des traductions modifiées.

Nous constatons donc que si l'on ajoute des données corrigées au corpus d'apprentissage, même en faible quantité, nous améliorons, de façon prévisible, la performance du système si l'on re-traduit ces même énnoncés (résultats sur le corpus PE dans le tableau III.2) mais on l'améliore également sur d'autres énnoncés (résultats sur le corpus $TEST$ dans le tableau III.2). Il serait cependant nécessaire de confirmer un tel résultat à l'aide d'un corpus plus important de post-éditions.

Poids $PE_{corr.}$	Enoncés corpus apprentissage	Corpus PE Enoncés $\neq$	Score BLEU	Corpus $TEST$ Enoncés $\neq$	Score BLEU
0	1 640 463	0 %	23,50	0 %	25,27
1	1 640 638 (+ 0,01%)	85 %	25,17	42 %	25,28
10	1 643 213 (+ 0,1%)	86 %	25,28	44 %	25,30
100	1 657 963 (+ 1,06%)	90 %	25,49	49 %	25,38
1000	1 815 463 (+ 11%)	90 %	25,73	65 %	25,51

TABLE III.2 – Résultats de l'ajout du corpus de traductions corrigées lors de l'apprentissage (systèmes sans ajustement des poids du modèle)

Exemples

Des exemples d'énoncés du corpus $TEST$ traduits avec le système *Baseline versus* le système *Baseline + 1000* PE_{corr} (c.-à-d. appris sur les données d'apprentissage standard augmentées de 1000 fois les données post-éditées) sont donnés dans la figure III.2. Nous remarquons, dans ces exemples, que le problème des mots inconnus persiste (ce qui n'est pas étonnant au vu de la taille du corpus post-édité – 175 énnoncés –) mais que, mis à part cela, les traductions données en exemple sont plus cohérentes avec l'énoncé source.

Enoncé source :	Au terme des échanges, la bourse de Prague bascula dans le négatif.
Baseline :	In terms of trade, the stock market in Prague in the negative *bascula*.
+ 1000 PE_{corr} :	At the end of trade, the Prague Stock Exchange *bascula* in the negative.
Enoncé source :	Paulson : le plan doit être efficace.
Baseline :	Paulson 's plan is to be effective.
+ 1000 PE_{corr} :	Paulson : the plan must be efficient.
Enoncé source :	On vous conseillera, comment choisir.
Baseline :	You can choose *conseillera*.
+ 1000 PE_{corr} :	We *conseillera*, how to choose.

FIGURE III.2 – Exemple d'énoncés du corpus $TEST$ traduites par le système *Baseline* et le système *Baseline + 1000* PE_{corr}

3.2 Ajustement des poids du système sur les post-éditions

Comme présenté dans le deuxième chapître de ce manuscrit, notre système de traduction de référence est fondé sur une modélisation log-linéaire composée de 14 modèles. La contribution de chaque modèle du système est estimée par une pondération et l'ensemble des pondérations des modèles constituent les paramètres ou poids du système.

Méthode d'ajustement des poids du modèle

Il a été montré que les performances des systèmes de traduction probabilistes peuvent être améliorées par l'ajustement des paramètres du modèle log-linéaire, notamment grâce à la stratégie *Minimum Error Rate Training* (ou *MERT*) qui règle les poids des modèles par minimisation d'un critère d'erreur sur un corpus de développement [Och 2003]. Ce procédé consiste à trouver, sur un corpus de développement, la combinaison de poids qui va permettre d'obtenir des traductions système aussi proches que possible des traductions données comme référence. En pratique, il n'est pas toujours vérifié que l'optimisation des poids sur le corpus de développement entraîne un gain de performance sur le corpus de test. Cela peut s'expliquer, en partie, par la distance entre les traductions produites par le système et les traductions *gold standard* habituellement données comme référence. L'idée est donc de remplacer, dans le corpus de développement, les traductions *gold standard* par les traductions du système de référence qui ont été corrigées lors de la post-édition.

Utilisation du corpus de post-éditions *versus* le corpus de traductions *gold standard*

La performance de référence de notre système de traduction *baseline* est obtenue avec des valeurs de poids du modèle définies par défaut dans la boîte à outils Moses.

Dans le cas de notre système de traduction de référence, l'ajustement des valeurs des poids sur le corpus de développement optimise bien le score BLEU sur ce corpus (score qui passe de 24,32 avant ajustement à 24,50 après ajustement), néanmoins, cette tendance n'est pas confirmée avec le corpus de test sur lequel les poids par défaut permettent d'obtenir un score BLEU de 25,27, contre un score BLEU de 25,05 avec les poids optimisés sur le corpus de développement. Le corpus de développement utilisé pour cet ajustement est décrit dans la partie 2.3 du chapître II. Il contient 1000 énoncés source et leurs traductions dites « de référence » faites par des professionnels.

Pour vérifier si l'origine des traductions données comme référence lors de l'ajustement des poids peut expliquer, en partie, l'échec constaté de l'ajustement des poids, nous avons utilisé le corpus PE pour ajuter les poids du système de référence en considérant comme traductions de référence les traductions *gold standard* (corpus PE_{std}) d'une part, et les traductions corrigées (corpus PE_{corr}) d'autre part.

Les différences entre les valeurs des poids issus de l'ajustement sur le corpus avec les *gold standard* (PE_{std}) et ceux issus de l'optimisation sur le corpus avec les traductions corrigées (PE_{corr}) sont données dans le tableau III.3.

L'analyse des différences entre les valeurs des poids issus de l'optimisation sur PE_{std} et ceux issus de l'optimisation sur PE_{corr} montre que :

- les poids liés au nombre de mots dans la phrase (+ 6 %) et au coût de réordonnancement des mots de la phrase (+ 7 %) sont plus importants pour le modèle optimisé sur les post-éditions ;
- il y a des variations notables des poids des 6 fonctions relatives au réordonnancement des mots de la phrase (de +/- 4 % à 11 %) entre les deux systèmes.

Ceci peut être interprété par le fait que les traductions post-éditées sont plus proches

Fonctions de trait	Poids ajustés avec $PE_{std.}$	Poids ajustés avec $PE_{corr.}$	Variation
$T(s/t)$	0,02	0,03	+ 0,01
$P_{lex}(s/t)$	0,08	0,05	- 0,03
$T(t/s)$	0,04	0,06	+ 0,02
$P_{lex}(t/s)$	0,03	0,04	+ 0,01
pp	0	0,03	+ 0,03
wp	-0,19	-0,25	**- 0,06**
$P(t)$	0,10	0,11	+ 0,01
d	0,05	0,12	**+ 0,07**
D_{mp}	0,04	0,08	+ 0,04
D_{mf}	0,05	0,01	- 0,04
$\mathbf{D_{sp}}$	0,08	0,01	**- 0,07**
$\mathbf{D_{sf}}$	0,12	0,04	**- 0,08**
$\mathbf{D_{dp}}$	0	0,09	**+ 0,09**
$\mathbf{D_{df}}$	0,13	0,02	**- 0,11**

TABLE III.3 – Différences entre les valeurs de poids issues de l'ajustement sur les *gold standard* (PE_{std}) *versus* sur les post-éditions (PE_{corr}) pour un même corpus de développement de 175 énoncés.

des traductions du système de référence que les traductions professionnelles *gold standard*. De ce fait, elles ont, globalement, le même nombre de mots (*cf.* modèles de pénalité sur le nombre de mots de la phrase) et l'ordre des mots est mieux conservé entre une phrase source et sa traduction (*cf.* modèles de ré-ordonnancement).

D'autre part, nous remarquons que le processus d'optimisation converge plus rapidement et plus efficacement avec l'utilisation des post-éditions. Comme le montre la figure III.3, le score BLEU passe de 33 à 66 en 9 itérations avec les post-éditions alors qu'il passe de 23 à 25 en 19 itérations avec les traductions professionnelles *gold standard*.

Evaluation automatique des systèmes selon la référence utilisée pour l'ajustement des poids de leurs modèles

L'objectif est de comparer la qualité des traductions obtenues par le système optimisé sur les traductions *gold standard* et celui optimisé sur les traductions post-éditées.

Pour le corpus PE, comme le montre la figure III.3, le score BLEU de 23,50 sans optimisation atteint un maximum de 24,90 avec l'optimisation sur les traductions *gold standard* alors qu'il augmente de 33,50 à 66,30 avec l'optimisation sur les traductions post-éditées. L'ajustement des poids avec les post-éditions est donc nettement le plus efficace, sur le corpus PE de post-édition, dans le sens où le système parvient à ajuster les poids du modèle de façon à obtenir des sorties du système très proches du corpus PE_{corr}.

Pour évaluer les résultats, les 1200 énoncés du corpus de test ont été traduits, d'une

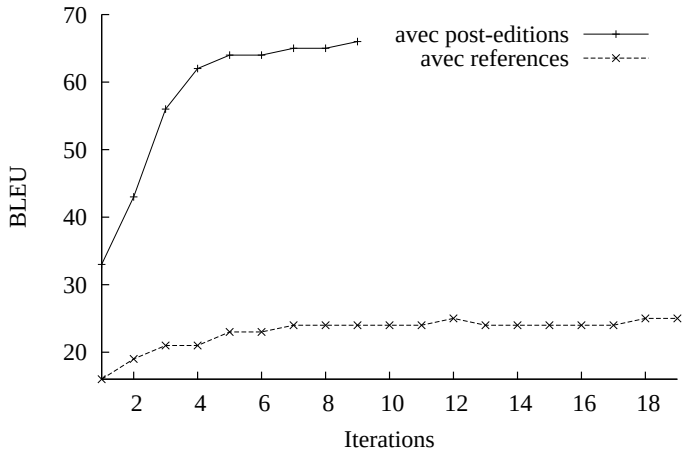


FIGURE III.3 – Evolution du score BLEU au cours de l'optimisation des poids avec MERT

	Sans ajustement des poids	Avec ajustement des poids sur PE_{corr}	sur PE_{std}
Score BLEU	25,27	25,36	25,43

TABLE III.4 – Scores BLEU, sur le corpus TEST, selon la référence utilisée pour l'ajustement des poids des fonctions de traits du modèle log-linéaire.

part, avec le système muni des poids optimisés sur les *gold standard* (PE_{std}) et, d'autre part, avec celui muni des poids optimisés sur les traductions corrigées (PE_{corr}). Bien que les deux modèles ne diffèrent que par leur combinaison de poids, au final, 65 % des hypothèses de traduction sont différentes entre les deux systèmes.

Les résultats présentés dans le tableau III.4 ne montrent aucune amélioration significative entre le système non optimisé et le système optimisé, et ce, quelque soit la référence utilisée pour l'optimisation. Ce résultat n'est pas surprenant car notre corpus de post-éditions possède l'inconvénient d'être relativement petit (175 énoncés).

Evaluation subjective des systèmes selon la référence utilisée pour l'ajustement des poids de leurs modèles

Afin de mieux juger les résultats obtenus, nous avons effectué une évaluation subjective au cours de laquelle trois évaluateurs volontaires ont comparé des hypothèses de traduction. Pour chaque énoncé à évaluer, nous avons demandé aux évaluateurs s'ils jugeaient meilleure l'hypothèse de traduction donnée par le système optimisé sur

PE_{corr}, celle donnée par le système optimisé sur PE_{std} ou s'ils les considéraient comme équivalentes. Les évaluateurs ont chacun évalué les 175 énoncés du corpus PE et 928 énoncés pris au hasard dans le corpus $TEST$, soit un total de 1103 énoncés.

Nous nous intéressons aux énoncés pour lesquels au moins deux des évaluateurs donnent le même jugement en procédant à un vote majoritaire. Cela représente environ 95 % des corpus PE et $TEST$ évalués. Les résultats sont donnés dans le tableau III.5. Nous pouvons y observer la même tendance que lors de l'évaluation automatique : pour le corpus PE les évaluateurs préfèrent majoritairement (à 64 %) l'optimisation sur les traductions corrigées (PE_{corr}) et pour le corpus $TEST$ les évaluateurs jugent, en majorité (dans 43 % des cas), équivalentes les deux optimisations.

Corpus	Nbre d'énoncés	Préférence des évaluateurs		
		$PE_{std.}$	$PE_{corr.}$	Indifférent
$TEST$	859 énoncés	33 %	24 %	**43 %**
PE	158 énoncés	16 %	**64 %**	19 %

TABLE III.5 – Préférence des évaluateurs selon la référence utilisée pour l'ajustement des poids du système de traduction.

3.3 Correction des résultats du système de traduction

Une autre approche consiste à utiliser le corpus annoté pour corriger les sorties du système. Nous avons choisi d'utiliser le principe de la post-édition automatique statistique.

Protocole expérimental

La tâche de post-édition est à l'origine manuelle : la traduction à corriger est éditée via une interface d'édition puis corrigée par un annotateur humain. Cette post-édition humaine peut toutefois être automatisée par des systèmes, qui, à l'image des systèmes de traduction automatique vont être entraînés à apprendre une correspondance entre les sorties d'un système et les corrections de ces sorties. Ce protocole appelé post-édition automatique statistique sera détaillé dans le chapitre VI de ce manuscrit.

Nous avons appris un système de traduction, sur le corpus PE, où les hypothèses de traduction proposées par le système de référence sont la langue source du système et leurs corrections effectuées par les annotateurs humains sont la langue cible. Ce système est nommé par la suite *post-éditeur automatique*. L'étape suivante consiste à appliquer ce système de post-édition automatique aux sorties du système sur le corpus PE et $TEST$ en procédant par décodage, comme avec un système de traduction standard qui traduit de la langue source (ici les hypothèses de traductions du système) vers la langue cible (ici les traductions corrigées).

Système	Corpus PE		Corpus $TEST$	
	énoncés $\neq$	Score BLEU	énoncés $\neq$	Score BLEU
Baseline	0%	23,50	0 %	25,27
+ post-éditeur	85%	24,58	40 %	24,32

TABLE III.6 – Résultats du post-éditeur automatique appris sur 175 post-édition manuelles

Résultats

Lorsque l'on post-édite automatiquement les traductions système du corpus $TEST$, le score BLEU passe de 25,27 à 24,32 (voir tableau III.6) avec 40 % de traductions corrigées. Une analyse qualitative des résultats nous permet de constater que la post-édition dégrade la qualité de la traduction du corpus $TEST$ avec des corrections faites à mauvais escient. Il faut noter que le post-éditeur automatique a été appris sur un corpus de 175 énoncés, ce qui est assurément insuffisant pour modéliser de façon efficace le comportement des post-éditeurs humains.

Sur le corpus PE, néanmoins, le post-éditeur automatique, qui corrrige 85 % des traductions, permet d'augmenter le score BLEU, de 23,50 à 24,58 (voir tableau III.6). Une brève analyse manuelle des résultats montre que le post-éditeur automatique permet bien, sur le corpus qui a servi à son apprentissage, de se rapprocher au plus près des post-éditions humaines. En effet, pour 71 % des énoncés, la post-édition automatique donne le même résultat que la post-édition humaine. La figure III.4 présente quelques exemples de ces énoncés.

Enoncé source :	Les lecteurs, par contre, ont l'avantage d'avoir une commande facile.
Baseline :	The readers, however, have the advantage of having a command easy.
+ post-éditeur :	The readers, however, have the advantage of having an easy command.
Enoncé source :	Le costume tyrolien bon marché semble atteindre son objectif.
Baseline :	The suit tyrolean cheap seems to achieve its goal.
+ post-éditeur :	The cheap tyrolean suit seems to achieve its goal.
Enoncé source :	Les pierres sont sales.
Baseline :	The stones are vile.
+ post-éditeur :	The stones are dirty

FIGURE III.4 – Exemple d'énoncé, du corpus PE, post-éditées automatiquement

4 Conclusion

Ces travaux présentent l'étude préliminaire de méthodes visant à améliorer un système de traduction automatique avec des rétro-actions d'utilisateurs. Ces expériences,

menées en utilisant un ensemble restreint de 175 énoncés post-édités, ont eu pour objectif d'expérimenter l'intégration des données à trois différents niveaux du processus de traduction : lors de l'apprentissage de la table de traduction du système, lors de l'ajustement des poids du modèle log-linéaire et sur les sorties de traduction du système. Pour cela, nous avons défini, expérimenté et évalué trois protocoles où les post-éditions manuelles sont utilisées pour :

- apprendre la table de traduction du système ;
- régler les poids du modèle de traduction ;
- et corriger les sorties du système.

Ces différentes expérimentations montrent que, les trois techniques permettent bien de corriger et améliorer le système de traduction pour les données déja rencontrées et corrigées (corpus PE) mais permettent difficilement de propager ces corrections sur de nouvelles données (corpus $TEST$). En effet, le corpus de post-éditions de 175 énoncés collecté est manifestement insuffisant pour nous permettre de généraliser les corrections sur de nouvelles données.

C'est pourquoi, le chapitre suivant présente la collecte d'un corpus de post-éditions manuelles à grande échelle (environ 12 000 énoncés). Ce corpus permettra l'analyse fine des corrections effectuées par les annotateurs sur les sorties de notre système de traduction automatique par le biais d'une étude quantitative et qualitative des post-éditions recueillies. Nous espérons, d'autre part, pouvoir confirmer et améliorer les résultats proposés ici grâce à des données plus importantes et développer, en sus, d'autres méthodes pour intégrer ces corrections au système.

Chapitre IV

Collecte d'un corpus de post-éditions

Le but de notre scénario applicatif est d'« apprendre » à partir de retours d'informations (ou *feedbacks* en anglais) d'utilisateurs représentés par des corrections de résultats du système de traduction. Le besoin pressenti pour nos expérimentations d'intégration des corrections de traductions est un corpus d'environ 10 000 traductions issues de notre système, revues et corrigées par des annotateurs humains.

1 La tâche de post-édition

1.1 Post-édition manuelle

La post-édition manuelle fait référence à une édition *a posteriori* des hypothèses produites par un système. Elle consiste en la révision humaine d'un texte traduit par une machine afin de lui donner du sens d'un point de vue grammatical et sémantique.

L'édition de l'hypothèse pouvant impliquer des modifications et des corrections d'erreurs, la tâche de post-édition peut donc être définie comme l'action de vérifier et corriger, si nécessaire, des textes produits par des systèmes de traduction automatique.

Le principe de post-édition tel que situé dans notre contexte est illustré dans la figure IV.1 : étant donnée une phrase source **S** en français et son hypothèse de traduction anglaise **T** fournie par notre système de référence, les annotateurs sont chargés de vérifier la qualité de l'hypothèse issue de notre système et de proposer une traduction vérifiée **T'** comportant, si nécessaire, des corrections de la phrase **T**.

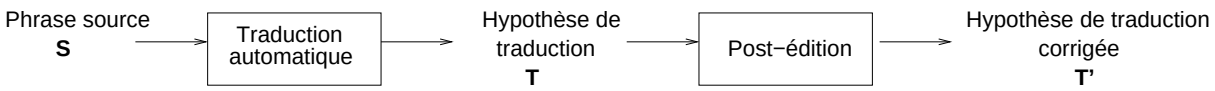


FIGURE IV.1 – Principe de post-édition de traductions automatiques

1.2 Corpus

Les données que nous avons choisi d'utiliser pour la collecte de post-éditions sont de même nature que les corpus journalistiques utilisés pour l'apprentissage et le test du système de référence (décrits dans la section 2.3 de la partie II). Le corpus est composé de 10 881 phrases issues de divers site Web journalistiques. Il convient de noter que le corpus utilisé pour la collecte de post-éditions est disjoint de celui utilisé pour l'apprentissage, développement et test du système de référence.

Lors d'expériences passées, nous avons eu l'occasion de constater, à plusieurs reprises, que les traductions de référence fournies avec les corpus bilingues alignés en phrases étaient souvent « éloignées » des traductions produites par les systèmes automatiques, de part leur caractère de traductions non-littérales. Le tableau 3.2 de l'annexe 2, présente des exemples de phrases extraites de notre corpus qui présentent cette caractéristique. Ces traductions sont faites, en amont, par des traducteurs professionnels, souvent indépendamment de tout système de traduction automatique et sans l'objectif d'être ré-utilisées par la suite, pour effectuer de l'apprentissage automatique de systèmes. Cela pose problème, entre autres, lors de leur utilisation pour juger de la qualité des hypothèses automatiques. Partant de ce constat, nous avons également fait post-éditer 1 500 traductions de référence fournies avec le corpus bilingue.

Nous avons donc sélectionné, pour la tâche de collecte de post-éditions, un corpus bilingue de 10 881 phrases dont nous avons fait post-éditer les 10 881 hypothèses de traduction issues de notre système et 1 500 traductions de référence de cet ensemble (extraits aléatoirement des 10 881 traductions de référence du corpus bilingue).

Le but de la tâche de post-édition est d'obtenir un ensemble de 10 881 quadruplets (items 1 à 4) et 1 500 quintuplets (items 1 à 5) contenant :

1. une phrase source (en français) ;
2. une hypothèse de traduction de notre système de référence (en anglais) ;
3. la correction de l'hypothèse de traduction (la post-édition humaine en anglais) ;
4. la traduction de référence fourni dans le corpus parallèle (en anglais) ;
5. la correction de la traduction de référence [28] (la post-édition humaine en anglais).

2 Méthode de collecte

La collecte d'une telle quantité de données nécessite : une plateforme de post-édition ; des annotateurs volontaires ; et un contrôle des annotations produites.

Construire un système complet pour la post-edition humaine est long et couteux à mettre en œuvre. Cela nécessite, entre autres, de créer une interface graphique, trouver, sélectionner et rémunérer les participants, et contrôler leurs contributions. Anfin de limiter les coûts relatifs à la collecte, nous avons choisi d'utiliser un outil de *crowdsourcing* qui permet d'organiser et faciliter la gestion de celle-ci.

28. Pour un sous-ensemble de 1 500 phrases du corpus, la post-édition de la traduction de référence permet d'obtenir un cinquième item qui s'ajoute aux quadruplets précédents.

2.1 Le *crowdsourcing*

Le terme *crowdsourcing* est un néologisme conçu en 2006, sa traduction littérale est « approvisionnement par la foule ». Le terme désigne un principe de fonctionnement qui consiste à utiliser la créativité, le savoir-faire et le temps disponible des internautes pour réaliser des tâches, résoudre des problème ou créer du contenu à moindre coût (la participation des internautes pouvant être rémunérée ou non).

Le *crowdsourcing* est mis en œuvre au moyen de plates-formes en ligne qui permettent à des internautes d'effectuer des tâches parcellisées.

Au jour d'aujourd'hui, de nombreux sites Web fonctionnent sur le modèle du *crowdsourcing* et proposent à des contributeurs volontaires de produire du contenu de façon collaborative.

Wikipedia [29], par exemple, est une célèbre encyclopédie en ligne qui permet à des contributeurs volontaires de participer à la rédaction des articles. La contribuation peut également se faire sous forme de jeu comme par exemple dans Games with Purpose [30] qui est une plateforme ludique dont les réponses des utilisateurs constituent des données destinées à l'apprentissage automatique. D'autres sites permettent d'héberger des créations personnelles puis de les revendre moyennant une commission (Fotolia [31], Wooshii [32]) ou encore de fournir des prestations en ligne en échange de micropaiement (Amazon Mechanical Turk [33], CrowdFlower [34], We Heart it [35] pour n'en citer que quelques unes).

Amazon Mechanical Turk (AMT) est une plateforme sur le Web qui permet de confier des tâches à des internautes en échange d'une éventuelle rénumération. Les internautes proposant leur services sont appelés « workers » et les tâches, qui sont appelées HITs pour « Human Intelligence Tasks », relèvent par exemple de l'étiquetage d'images, de l'analyse audio ou encore de la recherche sur Internet. En vue de notre collecte, nous avons opté pour la plateforme AMT pour collecter des annotations provenant d'une grande quantité d'annotateurs anglophones rémunérés mais non nécessairement experts de la tâche.

De nombreux travaux de recherche ont étudié l'efficacité des outils de *crowdsourcing* pour créer des données annotées pour le traitement automatique du langage naturel. Le faible coût de la main d'œuvre disponible sur de telles plates-formes ouvre de nouvelles possibilités pour l'annotation du texte et du discours, et a le potentiel de modifier la façon de générer des données pour les technologies de la langue.

2.2 Charte d'utilisation

L'usage des outils de *crowdsourcing* soulève de vives controverses et les problèmes légaux, économiques et éthiques impliqués par la méthode de collecte sont sujets à de

29. http ://fr.wikipedia.org/
30. http ://www.gwap.com
31. http ://fr.fotolia.com/
32. http ://wooshii.com
33. http ://www.mturk.com
34. http ://crowdflowers.com
35. http ://weheartit.com/tag/crowdsourcing

vastes débats [Fort et al. 2011].

En ce qui nous concerne, nous avons choisi, à l'image de ce qui a été fait dans [Gelas et al. 2011], de définir et respecter des principes d'utilisation de l'outil. La charte d'utilisation proposée comporte les lignes de conduite suivantes :

1. les informations collectées doivent être utilisées à but non lucratif et mises à la disposition de la communauté scientifique ;
2. chaque participant doit nécessairement être renseigné sur le contexte de la tâche : « Quel est l'organisme qui propose cette tâche ? », « Quel est son but ? » et « A quoi vont servir les données collectées ? » ;
3. la rétribution proposée doit être appropriée à la tâche et pouvoir justifier d'une rémunération à l'heure décente ;
4. les pays de provenance des participants sont restreints afin de réduire le plus possible les situations dans lesquelles les internautes considèrent le *crowdsourcing* comme leur principale source de revenus.

2.3 Problème des annotateurs non experts

La collecte de données *via* les outils de *crowdsourcing* pose cependant le problème de la fiabilité du contenu généré par des annotateurs non experts. En effet, cette approche est utilisée ici pour produire de grandes quantités de données d'apprentissage de référence que l'on présuppose « correctes ».

Or, les annotations sont souvent de qualité variable et pas toujours fiables, pour différentes raisons et ce, même si elles proviennent d'un expert humain. En premier lieu certains exemples sont difficiles à annoter et, de plus, les annotateurs peuvent être distraits ou fatigués au fur et à mesure qu'ils progressent dans la tâche.

Il est donc presque toujours nécessaire de procéder à un contrôle des annotations collectées.

Ce qui différencie les approches par *crowdsourcing* est qu'elles consistent à confier un travail à un groupe de gens ou à une communauté entière, au lieu de confier ce travail à une personne dédiée ou à un prestataire. Le problème est aussi que la communauté comprend forcément des internautes mal-intentionnés qui font appel à des stratégies frauduleuses pour obtenir une rétribution sans avoir fait la tâche demandée.

Fort de ce constat, Amazon Mechanical Turk propose des outils permettant de modéliser le comportement des annotateurs afin de détecter les participants mal-intentionnés et réduire ainsi ce problème.

3 Mise en œuvre de la collecte

Il s'agit ici de collecter des post-éditions de traductions anglaises d'énoncés francais du domaine « général », *via* une plateforme de collecte en ligne.

3.1 Interface

Nous avons choisi de découper le corpus à post-éditer en énoncés. Un énoncé est, dans la grande majorité des cas, une phrase mais peut aussi être une suite de quelques mots (ne représentant pas forcément une phrase au sens grammatical du terme — un titre par exemple—) ou une suite de deux ou trois phrases. L'interface développée pour la tâche (représentée dans la figue IV.2) considère le corpus comme une suite d'énoncés à post-éditer.

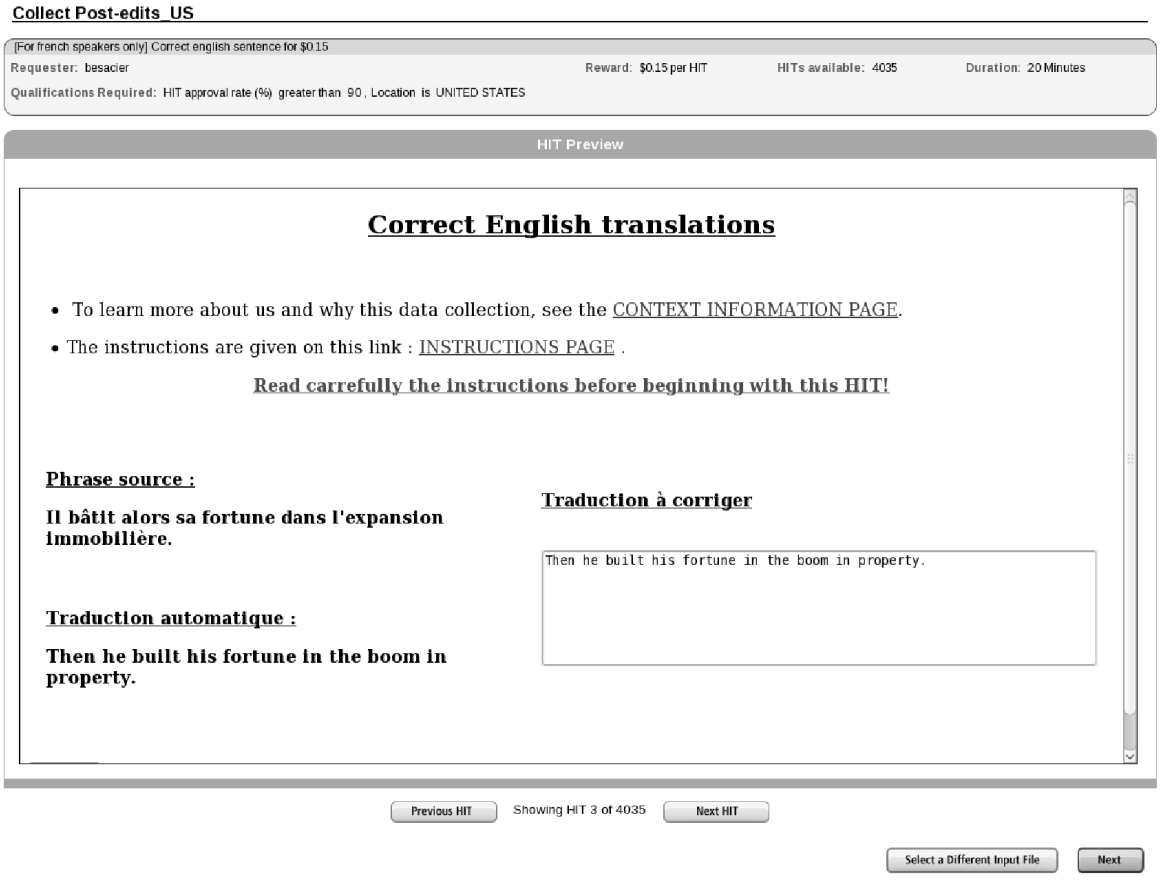
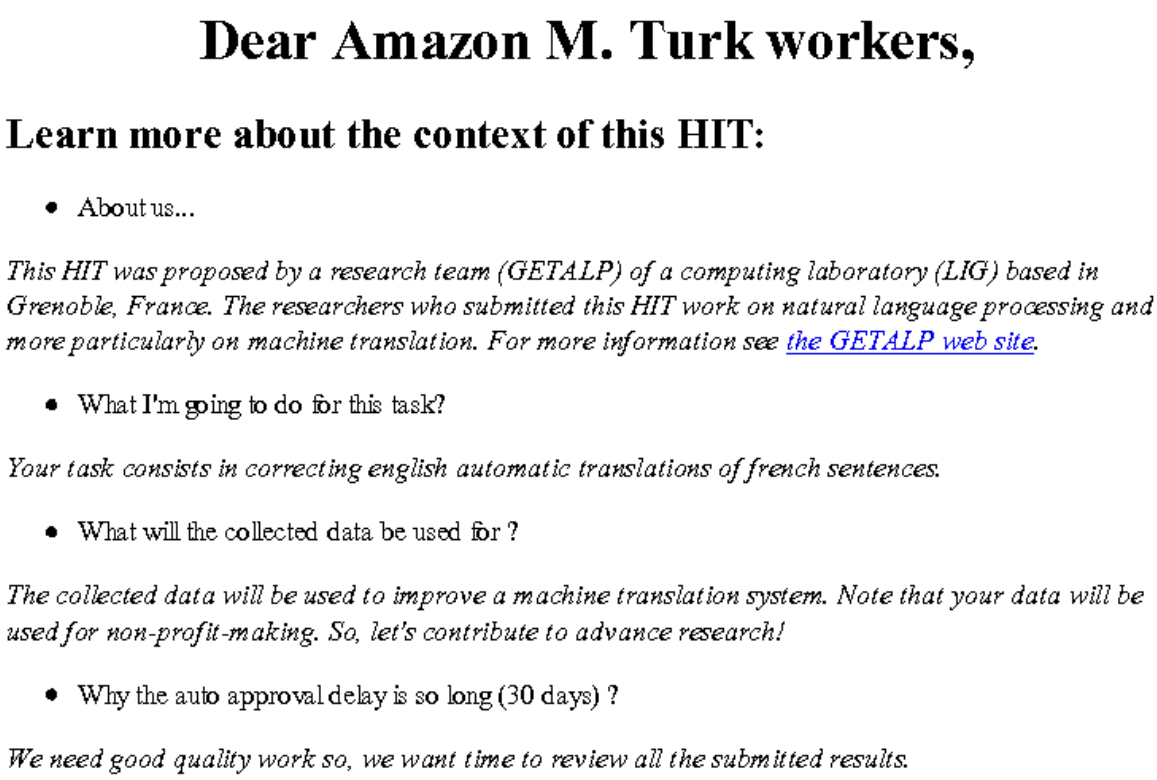


FIGURE IV.2 – Interface de la tâche de collecte de post-éditions

Afin de respecter la charte d'utilisation d'Amazon Mechanical Turk que nous avons définie dans la section 2.2, nous avons choisi de :

1. limiter les pays de domiciliation des participants à la France, aux Etats-Unis et au Canada ;
2. proposer une rémunération horaire d'environ 15 dollars (soit 0,15 dollar pour un HIT) ;
3. rediriger les participants à la tâche vers une page Web les renseignant sur le contexte de celle-ci (figure IV.3).



Dear Amazon M. Turk workers,

Learn more about the context of this HIT:

- About us...

This HIT was proposed by a research team (GETALP) of a computing laboratory (LIG) based in Grenoble, France. The researchers who submitted this HIT work on natural language processing and more particularly on machine translation. For more information see the GETALP web site.

- What I'm going to do for this task?

Your task consists in correcting english automatic translations of french sentences.

- What will the collected data be used for ?

The collected data will be used to improve a machine translation system. Note that your data will be used for non-profit-making. So, let's contribute to advance research!

- Why the auto approval delay is so long (30 days) ?

We need good quality work so, we want time to review all the submitted results.

FIGURE IV.3 – Contexte de la tâche de collecte de post-éditions

3.2 Profil des participants

Les participants à la tâche doivent nécessairement comprendre aisément la langue francaise et parler couramment l'anglais. Les participants n'ont pas à être nécessairement des anglophones natifs. L'interface de l'outil de *crowdsourcing* AMT, disponible en anglais uniquement, encourage une forte participation de locuteurs anglophones. Afin de nous assurer que les participants disposent d'une bonne compréhension de la langue française, nous avons rédigé les consignes de la tâche en français. Enfin, nous avons filtré les participants en tenant compte de leur pays de domiciliation (indiqué lors de leur inscription) et autorisé la participation seulement à ceux ayant indiqué résider aux Etats-Unis, au Canada ou en France.

3.3 Instructions de la tâche

La tâche consiste à corriger des traductions anglaises de phrases francaises. La figure IV.4 présente les instructions de la tâche, telles qu'elles ont été données aux participants. Il est indiqué que chaque traduction doit être vérifiée et, si besoin est, corrigée de façon à ce qu'elle représente une traduction correcte de la phrase française. On considèrera qu'une traduction est correcte si elle respecte les caractéristiques listées

ci-après :

1. la phrase anglaise produite doit être syntaxiquement et grammaticalement acceptable ;
2. tous les concepts/notions de la phrase source doivent être présents dans la traduction anglaise et vice versa (la traduction anglaise ne doit pas contenir de concepts/notions ne figurant pas dans la phrase source) ;
3. les modifications effectuées sur la traduction doivent être minimales. La traduction produite doit être la plus proche possible de la phrase française (respecter, dans la mesure du possible, l'ordre des mots, les tournures et le vocabulaire) tout en respectant les consignes données en 1. et 2. ;
4. la ponctuation doit être re-transcrite telle qu'elle est dans la phrase source française.

Il faut noter que les orthographes alternatives (Britaniques, Américaines) et les contractions sont autorisées.

3.4 Contrôle des annotations collectées

Le contrôle des annotations collectées concerne deux points particuliers. Il faut, d'une part, détecter le non-respect des consignes données et, d'autre part, détecter les participants mal-intentionnés (qui ne correspondent pas au profil demandé et/ou qui fraudent en éludant la tâche de post-édition).

Une automatisation du processus de contrôle des post-éditions étant difficilement envisageable, nous avons réalisé cette vérification manuellement : toutes les annotations soumises par les participants ont été vérifiées une à une et ont fait l'objet d'un contrôle strict, réalisé par nos soins. Toute post-edition ne vérifiant pas une ou plusieurs des quatre instructions données dans la consigne de la tâche a systématiquement fait l'objet d'un rejet. Un énoncé dont la post-édition a été rejetée est automatiquement ré-injecté parmi les énoncés à post-éditer et ce processus itère jusqu'à ce que la post-édition soumise soit acceptée.

Nous avons défini les participants « fraudeurs » comme des participants fautifs (qui ne respectent pas les consignes de la tâche) réguliers et récidivistes. Les fraudeurs identifiés se voient immédiatement bloquer l'accès à la tâche. Les fraudes les plus régulièrement constatées sont :

- la soumission systématique de l'hypothèse de traduction sans y apporter les corrections nécessaires ;
- la soumission d'une traduction de la phrase source issue d'outils traductifs automatiques, avec ou sans post-édition de cette traduction mais sans considération de l'hypothèse de traduction de notre système.

Alors que la première fraude est facilement et rapidement identifiable, la deuxième est plus difficile à déterminer en raison des nombreux systèmes de traduction automatique susceptibles d'être utilisés. L'analyse manuelle des résultats nous a permis de constater

Dear Amazon M. Turk workers,

Task Instructions

You must fairly understand french language and very well english language to contribute to this work. So, for this reason, the instructions will be given in french...

Votre tâche est de corriger des traductions anglaises de phrases francaises. Pour cela, il est nécessaire de comprendre aisément la langue francaise et de parler couramment anglais. Chaque traduction doit être vérifiée et, si besoin est, corrigée de façon à ce qu'elle représente une traduction correcte de la phrase française. On considèrera qu'une traduction est correcte si elle est syntaxiquement acceptable et si elle retransmet le sens de la phrase française.

Il est important de n'effectuer que les corrections nécessaires, c'est à dire de minimiser les modifications.

Veillez à ce que la ponctuation et les majuscules soient fidèlement retranscrites.

La phrase source représente la phrase francaise initiale. La traduction automatique représente la phrase source traduite en anglais par le traducteur automatique. Le champ "Traduction à corriger" contient, dans un premier temps, la traduction automatique qu'il vous faudra corriger.

1. **Lire attentivement** la phrase source et la traduction automatique
2. **Effectuer**, si nécessaire, **les corrections de la traduction** automatique dans le champ "Traduction à corriger" de façon à obtenir une traduction correcte de la phrase source
3. **Valider** avec le bouton "Submit" en bas de la page

Veillez, pour chaque phrase à corriger (ou HIT), à respecter les consignes suivantes :

- **minimiser les modifications : c'est à dire n'effectuer que les corrections nécessaires afin d'obtenir une phrase correcte "sémantiquement" et syntaxiquement**
- **la traduction doit être la plus proche possible de la phrase française (respecter, dans la mesure du possible, l'ordre des mots, les tournures et le vocabulaire)**
- **re-transcrire la ponctuation telle qu'elle est dans la phrase française (et ne pas vous fier à la ponctuation de la traduction proposée qui est souvent erronée)**
- **il est nécessaire que tous les concepts/notions de la phrase française soient traduits dans la phrase anglaise et *vice versa* (qu'il n'y ait pas de concepts/notions dans la traduction qui ne figurent pas dans la phrase source)**

Tout non respect de l'une de ces instruction pourra faire l'objet du rejet du HIT.

ATTENTION : l'utilisation de ***Google translate*** (ou tout autre traducteur automatique) pour produire directement les corrections est interdite et sera automatiquement sanctionnée !

Exemple :

Phrase source :

Voici un exemple de phrase qui est peut être mal traduite par le traducteur automatique.

Traduction automatique :

An example of sentence here is that is can be badly translated by the automatic translator.

Traduction à corriger

Here an example of sentence ~~here is~~ that ~~is can~~ **may** be badly translated by the automatic translator.

FIGURE IV.4 – Instructions de la tâche de post-édition

que l'outil largement utilisé dans ce but était *Google Translate*[36]. Nous proposerons, par la suite (dans la partie 4.2.0), une heuristique permettant d'évaluer *a posteriori* la

36. http ://translate.google.com

proportion de phrases soupçonnées de provenir du traducteur en ligne *Google Translate*.

4 Analyse du corpus de post-éditions collecté

4.1 Caractéristiques de la collecte

Durée Nous avons proposé un total de 12 381 énoncés à post-éditer : 10 881 hypothèses de traduction en anglais issues de notre système de traduction et 1 500 traductions professionnelles dites « de référence ». La collecte, qui a débuté le 8 décembre 2010 pour s'achever le 15 avril 2011, a été réalisée sur une durée de 4 mois et 12 jours. La durée passée par chaque internaute pour chaque post-édition (calculée comme l'intervalle de temps entre l'acceptation du HIT par un participant et sa soumission) n'est pas exploitable en raison d'un manque de maitrîse des différents composants de l'interface graphique proposée par Amazon Mechanical Turk. Il était, par exemple, possible de post-éditer l'énoncé avant d'avoir accepté le HIT.

Coût monétaire Les participants ont été rémunérés 0,15 dollars par énoncé corrigé, un total de 1 855 dollars ($12\,381 \times 0,15$) a donc été alloué au paiement des participants à la tâche. En sus, Amazon Mechanical Turk facture 10 % de cette somme (soit 185 dollars) pour les frais d'utilisation de l'interface. La collecte des 12 381 post-éditions aura donc coûté en tout 2040 dollars.

Nombre de participants Au total, 553 personnes inscrites sur Amazon Mechanical Turk ont participé à la tâche de post-édition. Parmi ces 553 participants, seuls 70 % d'entre eux (soit 396) ont réellement contribué à la tâche, c'est-à-dire on eu au moins une de leurs post-éditions validée. Autrement dit, 30 % des personnes ayant participé à la tâche ont vu tout leur travail rejeté pour cause de non respect des consignes. Par la suite, on différenciera un participant qui est une personne inscrite sur le site Amazon Mechanical Turk ayant soumis au moins un énoncé pour notre tâche de post-édition et un contributeur qui est un participant dont au moins un de ces énoncés post-édités a été accepté (les participants non contributeurs étant considérés comme des « fraudeurs »).

Validation des post-éditions Les statistiques réalisées *a posteriori* montrent qu'en moyenne, une phrase soumise avait 63 % de chances (soit environ deux chances sur trois) d'être rejetée pour cause de non respect des consignes (ou provenance de comportement frauduleux). La figure IV.5 représente le nombre de phrases validées en fonction du nombre de rejets précédant leur validation. On constate que 64 % des énoncés validés l'ont été lors de leur première soumission mais que les phrases restantes ont nécessité 2, 3 voire 4 soumissions et plus, pour certaines, avant d'être validées. Ces dernières sont des phrases que l'on peut considérer comme difficiles à post-éditer (des exemples sont donnés dans le tableau 3.3 de l'annexe 2).

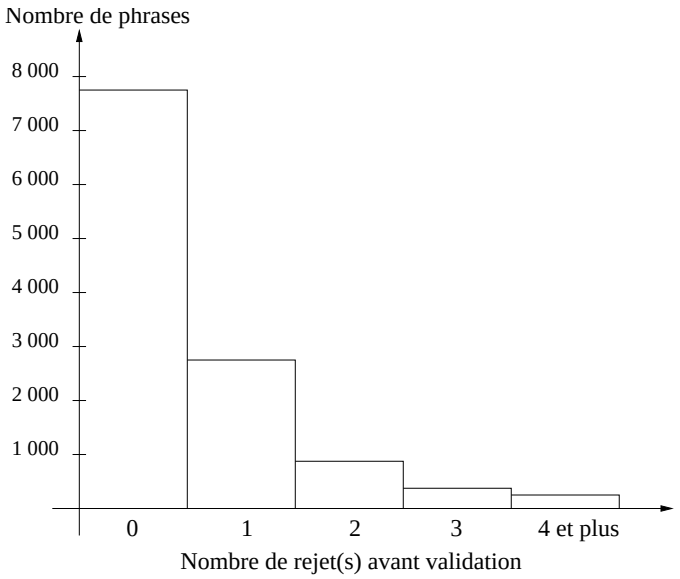


FIGURE IV.5 – Nombre de phrases soumises en fonction du nombre de rejets précédant leur validation

Contribution des participants En moyenne, un participant a post-édité 60 énoncés. Cependant, comme le montre le tableau IV.1, il existe de grandes disparités selon les participants si l'on considère le nombres d'énoncés post-édités et la contribution ou non dudit participant. Pour rappel, les participants « contributeurs » sont ceux dont au moins une post-édition a été acceptée alors que les participants « non contributeurs » sont ceux dont toutes les tentatives de contribution ont été rejetées. Nous pouvons remarquer que la moitié des participants (soit 52 % du total) sont des contributeurs ayant soumis moins de 10 post-éditions et que 6 participants (soit 1 % du total) ont contribué de façon significative en post-éditant plus de 500 phrases.

Nbre participants	Nombre de post-éditions ($= x$)				
	$x \leq 10$	$10 < x \leq 50$	$50 < x \leq 100$	$10 < x \leq 500$	$500 < x$
Contributeurs	293 (52%)	65 (11%)	12 (2%)	14 (2%)	6 (1%)
Non contributeurs	58 (10%)	49 (9%)	21 (4%)	32 (6%)	12 (2%)
Total	351 (62%)	114 (20%)	33 (6%)	46 (8%)	18 (3%)

TABLE IV.1 – Répartition des participants selon leur contribution

4.2 Evaluation de la qualité des post-éditions

Sous-ensemble d'évaluation

L'évaluation fine de la qualité des énoncés collectés est effectuée sur un sous-ensemble de 311 phrases du corpus post-édité *via* Amazon Mechanical Turk. Consi-

dérant l'ensemble des 12 381 phrases validées lors de la collecte comme répondant aux consignes de post-édition données, nous avons extrait les phrases à évaluer selon le protocole ci-après. Soit l'ensemble $\mathscr{P}$ des 12 381 post-éditions collectées :

- calcul du score TER moyen entre les post-éditions et leurs hypothèses de traduction sur $\mathscr{P}$. $\overline{TER}$ représente la distance moyenne d'édition (en terme de score TER) calculée entre les post-éditions et les hypothèses de traduction de l'ensemble $\mathscr{P}$;
- sélection des $\mathscr{P}_1$ post-éditions dont la distance à l'hypothèse de traduction (exprimée en score TER) se situe dans l'intervalle $[\overline{TER} - 4; \overline{TER} + 4]$: ce critère permet de sélectionner des phrases qui présentent un nombre d'opérations d'édition ni trop grand, ni trop petit ;
- parmi l'ensemble $\mathscr{P}_1$, sélection de $\mathscr{P}_2$ post-éditions selon les restrictions suivantes sur les post-éditeurs (ce critère permet d'obtenir un échantillon de post-éditions provenant de divers post-éditeurs) :
  - sélection d'au plus 2 post-éditions faites par des post-éditeurs ayant post-édité un total de moins de 10 phrases ;
  - sélection de 3 post-éditions pour les post-éditeurs ayant post-édité au total plus de 10 phrases.
- sélection aléatoire de 311 post-éditions parmi l'ensemble $\mathscr{P}_2$.

Détection des post-éditions d'origine frauduleuse

En plus du contrôle des tentatives de fraudes réalisé au fur et à mesure de la collecte, nous avons voulu évaluer le taux de suspicion de fraude ayant échappé à ce premier contrôle, *a posteriori*, sur l'échantillon de 311 énnoncés sélectionnés. Nous avons considéré la fraude consistant à utiliser le traducteur automatique *Google Translate* pour générer la post-edition (soit en utilisant directement sa sortie soit en la post-éditant). Nous avons défini et implémenté une heuristique prenant en compte le nombre de mots de la phrase source, la distance d'édition entre la post-édition et la traduction automatique de la phrase source donnée par Google pour classifier les post-éditions en quatre niveaux de suspicion. L'heuristique de classification est donnée dans l'annexe 2 et le résultat est donné dans le tableau IV.2. Seulement 3,5% des 311 phrases validées de l'échantillon d'évaluation sont sérieusement suspectées d'avoir été post-éditées en tenant compte du résultat de traduction de *Google Translate*.

Niveau de suspicion	aucun	faible	moyen	important	total
Nombre de phrases	273 (87,8%)	12 (3,8%)	15 (4,8%)	11 (3,5%)	311 (100%)

TABLE IV.2 – Suspicion d'utilisation de *Google Translate* lors de la post-édition

Evaluation humaine de la qualité des post-éditions

Le but est ici d'estimer la qualité des post-éditions collectées. Pour cela, un locuteur bilingue anglais/français, enseignant d'anglais de profession (et traducteur profession-

nel ayant pratiqué la post-édition) a jugé le sous-ensemble de 311 post-éditions de notre corpus.

Caractère *correctif* de la post-édition Dans un premier temps, chacune des 311 post-éditions a été soumise à un jugement contrastif avec l'hypothèse de traduction dont elle est issue. Une post-édition est ainsi jugée comme :

- **Correctrice** : si la post-édition améliore l'hypothèse de traduction (et ce même si elle contient encore des fautes) ;
- **Equivalente** : si la post-édition n'améliore pas l'hypothèse de traduction et que celle-ci comporte des erreurs qui auraient dû être corrigées dans la post-édition ;
- **Dégradante** : si la post-édition dégrade l'hypothèse de traduction (si elle introduit des erreurs qui n'étaient pas présentes dans l'hypothèse de traduction).

Caractère *traductif* de la post-édition De la même façon, indépendemment de son caractère correctif, chaque post-édition a été évaluée en tant que traduction de la phrase source ayant permis de produire l'hypothèse de traduction. Une post-édition est jugée selon qu'elle contient ou non des erreurs de traduction et la gravité des erreurs qu'elle contient :

- **Bonne** : si la post-édition ne comporte aucune erreur.
- **Acceptable** : si la post-édition comprend des erreurs mineures : faute d'inattention, majuscules oubliées/inappropriées, ou une ponctuation non adéquate.
- **Imprécise** : si le contenu de la post-édition est correct mais qu'il manque de précision dans la traduction : terme ou tournure inadapté, faute grammaticale, ou si le sens d'un mot français n'est pas vraiment rendu dans sa traduction (s'il existe un autre terme plus adapté).
- **Erronée** : si la traduction ne transmet pas, ou que partiellement, le contenu de la phrase source : omission de concept, contre-sens, faute de temps ou oubli majeur.

Ainsi, une post-édition contenant une erreur, même mineure et/ou localisée sur un seul mot, ne peut être considérée comme « Bonne » même si celle-ci n'affecte pas le reste de la phrase et, d'autre part, une post-édition contenant plusieurs erreurs est systématiquement rangée dans la classe correspondante à son erreur la plus grave.

Résultats Les résultats croisés de ces deux évaluations sont fournis dans le tableau IV.3. Ceux-ci montrent que 87,13 % des post-éditions améliorent l'hypothèse de traduction. Parmi les post-éditions collectées, 81,35 % (63,99 % + 17,36 %) sont des traductions irréprochables ou acceptables de la phrase source et 2,57 % contiennent une ou plusieurs erreurs qui font que la traduction reste fausse ou incomplète. On remarquera que ces dernières améliorent malgré tout l'hypothèse de traduction dont elles sont issues. Des exemples de phrases annotées de leurs jugements sont présentés dans le tableau 3.4 de l'annexe 2.

Post-édition	Bonne	Acceptable	Imprécise	Erronée	Total
Correctrice	189 (60,77%)	33 (10,61%)	41 (13,50%)	8 (2,57%)	271 (87,13%)
Equivalente	10 (3,21%)	20 (6,43%)	9 (2,89%)	0	39 (12,54%)
Dégradante	0	1(0,32%)	0	0	1 (0,32%)
Total (%)	199 (63,99%)	54 (17,36%)	51 (16,4%)	8 (2,57%)	311 (100%)

TABLE IV.3 – Jugements humains des post-éditions collectées

Comparaison avec des post-éditions professionnelles

Nous avons également évalué la qualité des post-éditions collectées via Amazon Mechanical Turk en comparant certaines d'entre elles avec des post-éditions faites par des traducteurs professionnels. La question à laquelle nous essayons de répondre par la suite est : *est-ce que des annotateurs bilingues non experts sont aussi efficaces que des traducteurs professionnels pour effectuer les corrections minimales d'une hypothèse de traduction donnée* ?

Post-éditions professionnelles Lors de travaux antérieurs, présentés dans [Specia 2011], un sous-corpus de 2 525 phrases sources inclues dans notre corpus de 10 881 phrases a été traduit par un système de traduction automatique basé sur les segments et les traductions obtenues ont été post-éditées par un traducteur professionnel. Le post-éditeur professionnel est un natif bilingue anglais/français avec un diplôme équivalent à une Licence (*First postgraduate degree*) en traduction. Les instructions données et le contexte de post-édition sont semblables aux notres : à partir de l'hypothèse de traduction, le post-éditeur est chargé d'effectuer les corrections minimales nécessaires pour aboutir à une traduction « publiable » de la phrase source. Le travail est payé et ne présente aucune contrainte de temps.

Comparaison des systèmes de traduction et résultats de post-éditions Avant de comparer les deux types de post-éditions, il est nécessaire de vérifier que les deux systèmes de traduction ayant produit les hypothèses de traduction sur lesquelles reposent les post-éditions, sont comparables. Par la suite nous désignerons par SMT_{ls} le système décrit dans [Specia 2011] et par SMT_{ref} notre système de traduction de référence décrit dans la partie II. Tout comme le notre, le système SMT_{ls} est un système de traduction basé sur les segments, entraîné à partir de la boîte à outils Moses. En terme de qualité, les deux systèmes sont très proches, sur le corpus de 2 525 phrases SMT_{ref} obtient un score BLEU de 20,20 alors que SMT_{ls} obtient un score BLEU de 20,76 (tableau IV.4). Sur les 2 525 phrases sources du corpus, 146 phrases présentent la même hypothèse de traduction quel que soit le système (SMT_{ref} ou SMT_{ls}). Autrement dit, dans 6 % des cas, les deux systèmes génèrent la même hypothèse de traduction.

Pour confirmer la proximité des deux systèmes, nous avons calculé pour chacun d'eux, la répartition des scores TER entre leurs hypothèses de traduction et les traductions de référence (on rappelle que les traductions de référence sont les mêmes pour les deux systèmes). Les résultats représentés dans la figure IV.6 montrent deux réparti-

tions très proches et illustre clairement que les deux systèmes présentent des résultats proches vis à vis des scores TER. Nous en concluons que les systèmes SMT_{ref} et SMT_{ls} sont très similaires.

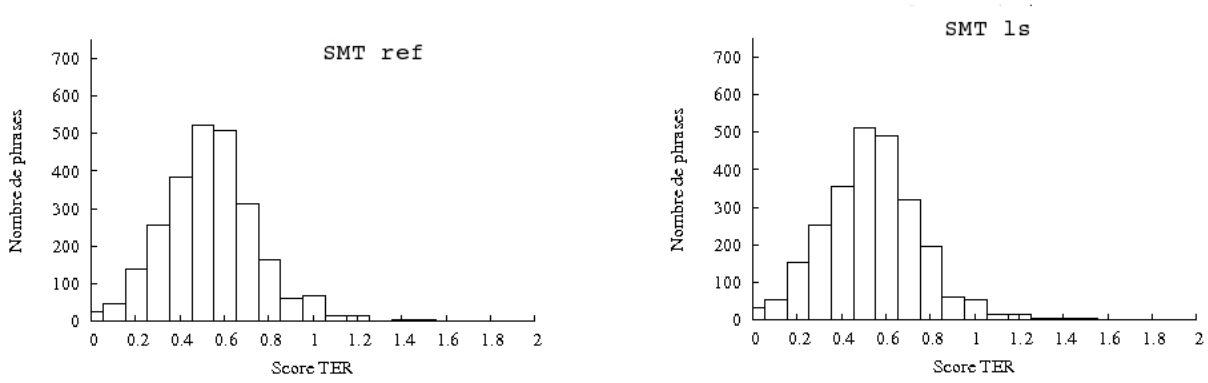


FIGURE IV.6 – Répartition des scores TER entre les hypothèses de traduction des deux systèmes (SMT_{ref} et SMT_{ls}) et les traductions de référence (sur 2 525 phrases).

La proximité entre les hypothèses de traduction des systèmes et leurs post-éditions, exprimée dans le tableau IV.4 en terme de score BLEU, nous laisse croire que les post-éditions collectées sur Amazon Mechanical Turk présentent moins de similitudes (calculées en considérant le recouvrement en n-grammes) avec les hypothèses du système SMT_{ref} que les post-éditions professionnelles avec les hypothèses du système SMT_{ls}. De la même façon, seulement 5,5 % des hypothèses de traductions n'ont pas été modifiées par les post-éditeurs d'Amazon Mechanical Turk contre 13 % pour les traducteurs professionnels. Pour 76 phrases (soit 3 % des cas) les deux post-éditions sont identiques.

Comparaison des post-éditions comparables Nous proposons maintenant de travailler sur les phrases du corpus pour lesquelles les post-éditions professionnelles et non-professionnelles sont directement comparables c'est-à-dire pour lesquelles les systèmes SMT_{ls} et SMT_{ref} ont fourni la même hypothèse de traduction.

Parmi les 2 525 phrases du corpus commum, 146 hypothèses de traduction sont identiques dans les deux contextes (post-édition professionnelle et post-édition non-professionnelle). Ces phrases sont de taille raisonnable pour permettre une étude puisqu'elles ont, en moyenne, 11 mots par phrase. Pour chacune des 146 phrases nous possédons donc le quintuplet suivant :

- la phrase source en français ;
- la traduction de référence (fournie par le corpus bilingue aligné) ;

	PE professionelles (avec SMT_{ls})	PE non-professionnelles (avec SMT_{ref})
Evaluation des SMT (score BLEU)	20,76	20,20
Taux de phrases non corrigées	13 %	5,5 %
Proximité entre hypothèses de traduction et post-éditions (score BLEU)	65,30	50,70
Hypothèses de traduction identiques entre SMT_{ls} et SMT_{ref}	146 (6 %)	
Post-éditions identiques entre SMT_{ls} et SMT_{ref}	76 (3 %)	

TABLE IV.4 – Comparaison des caractéristiques des systèmes SMT_{ls} et SMT_{ref} et des résultats de la post-édition (PE) selon la qualification professionelle des post-éditeurs : PE professionelles pour SMT_{ls} *vs* PE non-professionelles pour SMT_{ref}.

- l'hypothèse de traduction des systèmes automatiques (identique pour les deux systèmes) ;
- la post-édition du traducteur professionnel ;
- la post-édition non-professionnelle collectée sur le site de travail collaboratif AMT ;

Sur ces 146 phrases, 35 (ou 24 %) post-éditions sont identiques qu'elles aient été faites par le traducteur professionnel ou par un post-éditeur d'Amazon Mechanical Turk . Dans la majorité de ces cas (31 sur 35), il s'agit d'hypothèses de traduction ayant été jugées comme ne nécessitant pas de correction (post-édition identique à l'hypothèse de traduction).

Nous obtenons donc un ensemble de 111 phrases pour lesquelles nous pouvons directement comparer une post-édition professionnelle et une post-édition non-professionnelle différente. Cette comparaison a fait l'objet d'une évaluation humaine réalisée par un évaluateur bilingue (identique à celui ayant conduit l'évaluation précédente). Des exemples de comparaisons de post-éditions professionnelles et non-professionnelles d'une même hypothèse de traduction sont présentés dans le tableau 3.5 de l'annexe 2. Les résultats présentés dans la table IV.5 montrent que dans 26,13 % des cas, la post-édition réalisée par les post-éditeurs d'AMT est considérée comme meilleure que la post-édition professionnelle. Nous pouvons noter que dans une grande majorité des cas (c'est-à-dire 67,6 % des phrases), les deux post-éditions sont jugées de qualités équivalentes. En d'autres termes, 93,7 % des post-éditions que nous avons collectées *via* Amazon Mechanical Turk sont considérées comme étant de qualité professionelle. Cette étude nous permet d'inférer que le corpus de post-éditions que nous avons collecté est de qualité au moins équivalente au corpus de post-éditions professionnelles.

Préférence	Nombre de phrases
Post-édition non-professionelle	29 (26,1 %)
Post-édition professionelle	7 (6,3 %)
Post-éditions équivalentes	75 (67,6 %)

TABLE IV.5 – Comparaison de qualité entre post-éditions professionnelles et non-professionnelles sur 111 phrases

4.3 Analyse des corpus

Par la suite nous nommerons « *corpus-10881* » le corpus de 10881 phrases contenant la post-édition de l'hypothèse de traduction et « *corpus-1500* » le corpus de 1500 phrases contenant la post-édition de l'hypothèse de traduction ainsi que la post-édition de la traduction de référence.

Analyse des post-éditions des hypothèses de traduction

La tâche de post-édition nous a fourni un corpus de 10 881 phrases (*corpus-10881*) pour analyser les corrections des hypothèses de traduction de notre système de référence. Des exemples de corrections d'hypothèses de traduction sont donnés dans la figure IV.6.

Phrases source	**Hypothèse de traduction**	**Hypothèse de traduction corrigée**
• La police anti-émeutes les ont aussitôt encerclés et sont intervenus sans ménagement, jetant plusieurs d'entre eux à terre. • Le cinquième candidat affirme ne soutenir ni le pouvoir, ni l'opposition. • Forte mobilisation à Copenhague et à travers le mode, pour le climat. • Il y a des rivières qui s'assèchent en Afrique, des cours d'eau où l'on peut marcher comme on ne l'avait jamais fait avant.	• The anti-riot police were immediately surrounded and spoke blunty, several of them on land. • The fifth candidate says it support nor the current leadership, nor the opposition. • Strong involvement in Copenhague and in the world climate. • There are rivers are drying up in Africa, rivers where you can walk as it had never done before.	• ~~The~~ **A**nti-riot police**men** ~~were~~ immediately surrounded **them** and ~~spoke blunty~~ **stepped in ruthlessly**, **throwing** several of them ~~on land~~ **to the ground**. • The fifth candidate says ~~it~~ **he** support ~~nor~~ **neither** the current leadership, nor the opposition. • Strong ~~involvement~~ **mobilization** in Copenhague and ~~in~~ **across** the world **for the** climate. • There are rivers ~~are~~ drying up in Africa, ~~rivers~~ **watercourses** where ~~you~~ **one** can walk as it had never done before.

TABLE IV.6 – Exemples de corrections faites sur les hypothèses de traductions de notre système de référence

Le ratio entre la taille des hypothèses de traduction et les post-éditions est de 1,02 ;

cela signifie que la post-édition ajoute en moyenne peu ou aucun mot(s) à l'hypothèse de traduction. D'autre part, 9 % des hypothèses de traduction ont été jugées par les post-éditeurs comme ne nécessitant pas de correction lors de la post-édition, c'est-à-dire que 9 % des résultats de notre système de traduction de référence sont considérés comme de parfaites traductions de la phrase source.

Analyses des post-éditions des traductions de référence

Le sous-corpus de 1500 phrases (*corpus-1500*), nous permet d'analyser, en plus des corrections des hypothèses de traduction, les corrections des traductions fournies comme références dans le corpus bilingue. Des exemples de corrections de traduction de référence sont donnés dans la figure IV.7.

Phrase source	**Traduction de référence**	**Traduction de référence corrigée**
• Mais une fiscalité insuffisante peut également produire les mêmes effets.	• Too little taxation can do the same.	• **But** ~~Too little~~ **an insufficient** taxation can **also** do ~~have~~ the same **effects**.
• Le malaise français n'a certainement pas été induit par ces réformes.	• The French malaise has nothing to do with any of them.	• The French malaise ~~has nothing to do with~~ **was certainly not induced by** a~~ny of them~~ **these reforms**.
• Mais quelle est la signification réelle de ces deux principes ?	• But what do solidarity and subsidiarity really mean ?	• But what ~~do solidarity and subsidiarity really mean~~ **is the real meaning of these two principles** ?
• Les traités européens expriment clairement cette subsidiarité verticale.	• In the European Treaties, we find a clear expression of vertical subsidiarity.	• ~~In~~ **T**he **e**uropean **t**reaties ~~we find a clear expression of~~ **clearly express this** vertical subsidiarity.

TABLE IV.7 – Exemples de corrections faites sur les traductions de référence professionnelles fournies dans les corpus paralèles bilingues

Le taux de phrases non-corrigées de ce sous-corpus est donné dans la tableau IV.8. Nous y observons la même tendance que celle constatée sur le corpus de 10 881 phrases (*corpus-10881*) : 10 % des hypothèses de traduction dérivées de notre système de référence ne nécessitent aucune correction lors de la post-édition. Néanmoins, les mêmes statistiques faites sur les traductions de référence montrent que seulement 28 % d'entre elles sont condidérées comme correctes, autrement dit, 72 % des traductions proposées comme référence dans les corpus bilingues ont nécessité une correction lors de la post-édition.

Ceci peut être expliqué par le fait que les traductions professionelles de référence fournies avec les corpus bilingues alignés sont, dans le cas des corpus utilisés ici, produites dans un contexte de traduction à l'échelle de textes : la traduction est effectuée

en considérant le texte dans son intégralité et non la phrase en tant qu'unité indépendante. En segmentant les textes ainsi traduits au niveau des phrases, nous pouvons aisément imaginer qu'un concept manquant dans une traduction peut en fait apparaître dans une phrase adjacente (entraînant, par la même occasion un concept superflu dans ladite phrase).

De telles traductions effectuées au niveau du texte se révèlent donc comme n'étant pas des plus appropriées dans un contexte d'apprentissage automatique phrase-à-phrase d'un système de traduction probabiliste.

Taux de traductions non corrigées	*corpus-10881*	*corpus-1500*
Hypothèse de traduction	9 %	10 %
Traductions de référence	-	28 %

TABLE IV.8 – Taux de traductions non corrigées lors de la post-édition (considérées comme correctes par les post-éditeurs) selon le type de traduction

Corpus	**Types de traduction comparées**	**TER**	**d**	**WER**
corpus-10881	Hypothèse & Référence	55,1	73,0	58,5
	Hypothèse & Hypothèse corrigée	24,3	38,6	26,3
	Hypothèse corrigée & Référence	51,7	67,8	54,4
corpus-1500	Hypothèse & Référence	53.8	70,5	56,0
	Hypothèse & Hypothèse corrigée	24,2	34,9	23,6
	Hypothèse corrigée & Référence	47,6	63,0	48,0
	Référence & Référence corrigée	21,9	28,1	20,9
	Hypothèse corrigée & Référence corrigée	30,9	56,6	40,1

TABLE IV.9 – Distance entre les différents types de traduction

Distance entre les différents types de traduction

L'objectif est ici de mesurer la similarité entre les différents types de traduction. Pour cela, nous distinguons l'hypothèse de traduction automatique du système probabiliste de réference (nommée par la suite *hypothèse*), des traductions « validées par l'humain » qui sont : l'hypothèse de traduction corrigée lors de la post-édition (nommée par la suite *hypothèse corrigée*), la traduction donnée comme référence dans les corpus alignés (nommée par la suite *référence*), et cette même traduction corrigée lors de la post-édition (nommée par la suite *référence corrigée*).

Pour mesurer la similarité entre deux chaînes de caractères nous utilisons : d'une part les scores WER et TER (définis dans la section 3.2) basés sur la distance d'édition de Leivenshtein, et d'autre part une métrique d basée sur le recouvrement en n-grammes, elle est inspirée du score BLEU (définit dans la section 3.2) et calculée par la formule $d = 100 - scoreBLEU$. Les métriques WER, TER et d peuvent être interprêtées comme des mesures de l'éloignement ou de la différence entre deux corpus composés de chaînes de caractères. Le résultat de ces trois métriques, appliquées aux différents types de traduction des corpus *corpus-1500* et *corpus-10881*, sont donnés dans le tableau IV.9. De la même façon, la figure 4.3 illustre la distance en termes de score TER calculé entre *hypothèses*, *hypothèses corrigées* et *références* sur le corpus de 10 881 phrases.

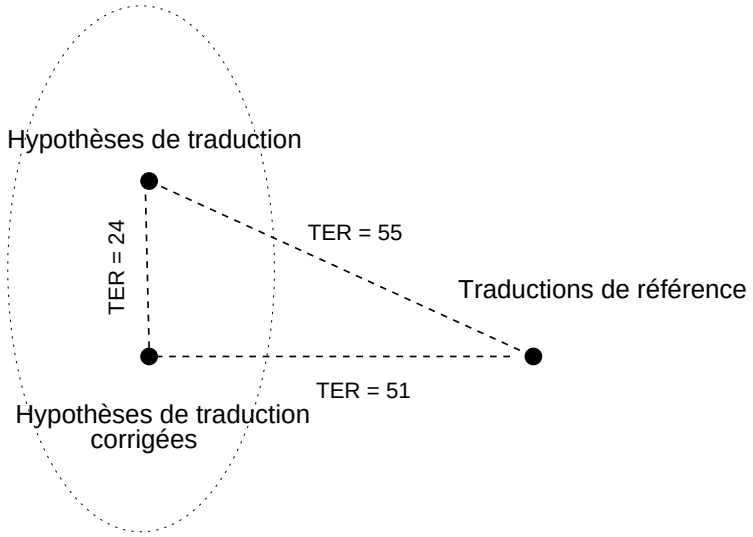


FIGURE IV.7 – Distances (en termes de score TER) entre les différents types de traduction sur le corpus de 10 881 phrases

Nous observons que la distance entre les traductions de référence et les hypothèses de traduction est deux fois plus importante que celle qui relie ces dernières à leurs corrections. De la même façon, les hypothèses de traductions corrigées et les traduction de référence sont excessivement éloignées alors que toutes deux sont censées être des traductions « correctes » des mêmes phrases sources. A titre indicatif, les répartitions des phrases des corpus en fonction des distances d'édition (Score WER calculé à l'échelle de la phrase) entre les traductions sont données dans l'annexe 2.

5 Conclusion

Nous avons collecté un corpus de 12 381 hypothèses de traduction post-éditées (environ 300 000 mots) soit l'équivalent d'environ 1200 pages de traducteur professionnel [37].

37. Dans la profession, on considère qu'une page de traducteur compte 250 mots

Une analyse subjective et comparative de ce corpus montre la qualité « professionnelle » des post-éditions collectées. La collecte ainsi que l'analyse de la ressource ont fait l'objet d'une publication dans la conférence internationale LREC [38] en mai 2012 [Potet et al. 2012] et le corpus est actuellement utilisé pour des travaux relatifs à l'estimation de confiance pour la traduction automatique au Laboratoire d'Informatique de Grenoble [Luong 2012].

Ce corpus est mis gratuitement à la disposition de la communauté scientifique et son téléchargement est possible, en ligne, à l'URL suivante :
`http://www-clips.imag.fr/geod/User/marion.potet/index.php?page=download`

38. Language Resources and Evaluation Conference

Chapitre V

Premières pistes explorées pour exploiter le corpus de post-éditions collecté

Ce chapitre décrit différentes expérimentations visant à exploiter le corpus de post-éditions précédemment collecté afin d'améliorer le système de traduction de référence décrit au chapitre II. Après avoir partitionné le corpus collecté en trois sous-ensembles (pour l'apprentissage, le développement et le test des systèmes), nous évaluons deux protocoles expérimentaux : l'enrichissement du système de traduction de référence par ajout d'une table de traduction apprise sur les post-éditions et la post-édition automatique des résultats du système de référence. Même si les pistes explorées se sont avérées *in fine* peu performantes, il nous a cependant semblé important de les détailler ici car elles ont contribuées à une meilleure maîtrise de la tâche et de la problématique.

1 Sous-ensembles pour l'apprentissage, le développement et le test

Pour les expérimentations qui vont suivre (dans ce chapitre et dans les suivants), le corpus de 10 881 post-éditions précédemment collecté (partie IV) est divisé en trois sous-ensembles : 8681 phrases pour l'apprentissage des systèmes, 1000 phrases pour le développement des systèmes et 1200 phrases pour le test des systèmes appris. Un récapitulatif des ces sous-ensembles est donné dans le tableau V.1.

Le corpus utilisé pour l'apprentissage des systèmes est de la même nature que les corpus journalistiques présentés dans la partie 2.3. Les corpus utilisés pour le développement et le test des systèmes sont constitués d'articles extraits du site Web français *Les Echos*[39], datant de mi-décembre 2009 et traduits en anglais par des professionnels de l'agence CEET[40] (Central and Eastern European Translations).

39. http ://www.lesechos.fr
40. http ://www.ceet.eu

Utilisation	Nature du corpus		Taille	
	Origine	Période	Nb. phrases	Nb. mots
Apprentissage	Divers sites Web	2007-2009	8 681	246 990
Développement	Site Web « Les echos »	mi-décembre 2009	1 000	26 846
Test	Site Web « Les echos »	mi-décembre 2009	1 200	32 968

TABLE V.1 – Description des corpus utilisés pour les expérimentations utilisant le corpus de post-éditions collecté dans la partie IV

2 Système enrichi d'une table de traduction apprise sur les post-éditions humaines

L'idée est ici d'enrichir le modèle de traduction du système de référence avec une table de traduction apprise sur les post-éditions collectées.

2.1 Protocole expérimental

Le principe consiste à combiner le modèle de traduction de référence avec une table de traduction additionnelle, apprise sur les données post-éditées.

La boite à outils Moses permet, de part son implémentation, l'utilisation de plusieurs tables de traduction. Cette technique est appelé « tables de traduction multiples » (ou *multiple translation tables* en anglais).

Le principe nécessite d'apprendre une table de traduction sur le corpus d'apprentissage de post-éditions puis d'ajouter cette table de traduction dans le modèle de traduction initial avant d'ajuster les poids du modèle à l'aide de la technique MERT. Après cette opération, le modèle de traduction est composé d'une combinaison coefficientée de la table de traduction du système de référence TT_{ref} et de la table de traduction apprises sur les post-éditions TT_{pe}. Chaque table de traduction possède alors ses propres ensembles de poids ajustés.

Les deux tables de traduction sont ensuite utilisées lors du décodage : les options de traduction sont collectées dans les deux tables et si une même option de traduction (en termes de segments source et cible identiques) est trouvée dans les deux tables de traductions, deux options de traduction distinctes sont créées pour chacune des occurences.

2.2 Résultats

Sur le corpus de test, 21 % des énoncés présentent des résultats de traduction différents selon s'ils sont traduis par le système de traduction de référence ou celui enrichi du modèle appris sur les post-éditions humaines.

Les résultats de l'évaluation du système de référence et du système enrichi sont donnés en terme de score BLEU sur le corpus de développement et de test dans le tableau V.2. Sur le corpus de test, le système de référence obtient un score BLEU

Corpus	Système de référence TT_{ref}	Système enrichi $TT_{ref} + TT_{pe}$
Developpement	24.32 (24.50)	20,71 (24,61)
Test	25,27 (25.05)	23,80 (23,78)

TABLE V.2 – Performances — en termes de score BLEU (avec les poids ajustés à l'aide de la technique MERT)— du système de référence (TT_{ref}) *versus* du système de traduction enrichi d'une table de traduction apprise sur les post-éditions humaines ($TT_{ref} + TT_{pe}$)

de 25,27 (avec les poids par défaut). Lorsque l'on intègre dans le système la table de traduction apprise sur le corpus de 10 881 post-éditions, ce score décroit à 23,80. L'ajustement des poids (entre autres ceux des deux tables de traduction TT_{ref} et TT_{pe}) sur le corpus de développement ne permet pas d'améliorer le score du système enrichi sur le corpus de test (Score BLEU de 23,80 avant ajustement et 23,78 après ajustement des poids).

Des exemples de phrases pour lesquelles le système enrichi génère des résultats de traduction de meilleure qualité que le système de référence sont donnés dans la figure V.3.

Phrase source	Système de référence TT_{ref}	Système enrichi $TT_{ref} + TT_{pe}$
Forte mobilisation à Copenhague et dans le monde pour le climat.	Strong *involvement* in Copenhague and in the *world climate*.	Strong *mobilisation* in Copenhague and in the *world for the climate*.
Avez-vous déjà essayé de jeter une bouteille plastique dans le centre ville ?	Have you *tried* to lay a *plastic bottle* in the downtown ?	Have you *already tried* to lay a *bottle plastic* in the downtown ?
Le cinquième candidat affirme ne soutenir ni le pouvoir actuel, ni l'opposition.	The fifth candidate says it support *nor* the current leardership, nor the opposition.	The fifth candidate says it support *neither* the current leardership, nor the opposition.
Si aucun candidat n'atteint ce seuil, un second tour doit être organisé dans les deux semaines.	If no candidate this threshold, a second round must be *held within two weeks ago*.	If no candidate *reached* this threshold, a second round must be *organised in the two weeks*.

TABLE V.3 – Exemples de comparaisons de traductions faites par le système de référence (TT_{ref}) *versus* du système de traduction enrichi d'une table de traduction apprise sur les post-éditions humaines ($TT_{ref} + TT_{pe}$)

3 Système de post-édition automatique appris sur les post-éditions humaines

L'idée est ici d'utiliser les post-éditions humaines préalablement collectées pour entraîner un système de post-éditions statistique à appliquer sur les résultats du système de traduction référence.

3.1 Système de post-édition probabiliste de référence

Le système de traduction de référence utilisé est celui présenté dans le chapitre II et le corpus de post-édition utilisé pour l'apprentissage et le test du SPES est celui présenté dans le chapitre IV.

Nous considérons les sous-ensembles précédemment définis dans la partie 1 : 8 681 phrases pour l'apprentissage du système de post-édition, 1 000 phrases pour le développement du système et 1 200 phrases pour le test du SPES appris.

Pour rappel, à chacune des 10 881 phrases sources françaises du corpus correspond une hypothèse de traduction en langue anglaise donnée par notre système de référence et deux traductions dites « de référence » : une post-édition de l'hypothèse de traduction du système de référence et une traduction professionnelle indépendante fournie avec le corpus parallèle.

Par la suite nous désignerons par SPES un Système de Post-Edition Statistique.

Apprentissage

Nous avons considéré la post-édition automatique comme une tâche de traduction exécutée par un système automatique à base de segments (PBMT) où le corpus source contient les sorties brutes du système et le corpus cible la version post-éditée (corrigée) de ces traductions brutes.

Notre système de post-édition présente la même architecture que notre système de traduction de référence et a été entrainé avec les mêmes outils (Moses, SriLM et GIZA++).

Le modèle de langage utilisé par le SPES, et conservé pour la suite des expérimentations de ce chapitre, est identique à celui de notre système de traduction de référence.

Table de traduction du modèle

L'apprentissage du système de post-édition statistique nous fournit, entre autres, une table de traduction où les segments traduits par notre PBMT de référence sont alignés avec leur post-édition humaine correspondante. A la manière d'un modèle de traduction automatique probabiliste, le SPES considère comme entrée la sortie brute du PBMT pour générer, à partir de la table de traduction, une nouvelle hypothèse de traduction.

Le résultat est une table de correspondance entre les segments corrigés lors de la post-édition et les segments d'origine (les hypothèses de traduction résultats du système

de traduction de référence). Des exemples d'alignements obtenus sont proposés dans la figure V.1.

1 : demanded of personal ||| asked for personal
2 : live from new experiences ||| to live new experiences
3 : survivante ||| survivor
4 : the city in the five lakes ||| the city with its five lakes
5 : the perspective of ||| the prospect of
6 : fairness ||| global equity
7 : a command easy ||| an easy command
8 : a decline in sales of 16 % ||| a decrease in sales of 16 %
9 : tsunami bavarian ||| bavarian tsunami
10 : troncs cerebral similar ||| similar brain stems
11 : cafards ||| cockroaches
12 : camera of moments of personality of ||| camera moments of personality of
13 : can be very ||| can appear very
14 : comparison nervous ||| nervous comparison
15 : electeurs free ||| free electors
16 : fêtera ||| celebrate
17 : i am spanish ||| i am a spaniard
18 : i live in rome for 25 ||| i have lived in rome for 25
19 : in the system of blood in quebec ||| in the blood system in quebec
20 : the opposition feels quite an improvement palpable ||| the opposition feels quite a palpable improvement

FIGURE V.1 – Extrait de la table de traduction apprise entre les hypothèses de traduction et leur correction

Cette table de correspondance nous permet de faire ressortir les différents types de corrections apportées par les annotateurs. Nous pouvons les classer en deux catégories : les comportements de type subsitution (où un segment m_1 est remplacé par un segment m_2, les segments pouvant être nuls $\emptyset$) et les comportements de type ré-ordonnancement (où un segment est présent à la fois dans l'hypothèse et sa correction mais sans occuper la même place).

Les exemples 7, 9, 14, 19 et 20 de la figure V.1 présentent des corrections de type ré-ordonnancement. La liste ci-après présente les corrections de type substitution :

$m_1 \rightarrow m_2$ remplacement d'un segment par un autre (exemples 1, 2, 4, 5, 6, 8, 13 et 17 de la figure V.1) ;

$m_1 \rightarrow m_2$ traduction d'un segment inconnu par le système (exemples 3, 10, 11, 15 et 16 de la figure V.1) ;

$\mathbf{m} \rightarrow \emptyset$ suppression d'un segment (exemple 12 de la figure V.1) ;

$\emptyset \rightarrow \mathbf{m}$ ajout d'un segment (exemple 18 de la figure V.1).

Evaluation du modèle

La qualité des post-éditions produites a été évaluée avec la métrique TER (Translation Error Rate) et le score BLEU. Pour rappel, le score TER reflète le nombre d'opérations (insertion, supression, susbstitution de mots et décalage de segments) nécessaires pour transformer une hypothèse de traduction en une traduction de référence alors que le score BLEU est une moyenne géométrique de la précision en n-grammes. Pour constater une bonne qualité de traduction, il faut chercher à minimiser le score TER et maximiser le score BLEU. Pour nous assurer que les différences constatées entre les scores sont significatives, nous avons calculé des tests statistiques de significativité des résultats obtenus en terme de score BLEU en utilisant la méthode de ré-échantillonnage (ou « bootstrap resampling method ») proposée par Philipp Koehn dans [Koehn 2004].

Système	Scores TER (*BLEU*)
PBMT	22.8 (*62.1*)
PBMT + SPES	24.1 (*60.0*)

TABLE V.4 – Performance du système de post-édition automatique statistique (SPES) de référence.

Les résultats obtenus par notre système de post-édition statistique de référence sont donnés dans le tableau V.4. Le système de traduction de référence (désigné par PBMT) obtient un score TER de 22,8 et un score BLEU de 62,1. Si l'on applique un système de post-édition automatique aux résultats de traduction de ce PBMT, on obtient alors un score TER de 24,1 et un score BLEU de 60,0. L'usage du post-éditeur automatique dégrade donc malheureusement significativement la qualité des sorties « brutes » du système de traduction de référence.

Les travaux suivants visent à tester différentes pistes d'amélioration afin de mieux exploiter le corpus collecté dans un contexte de post-édition statistique.

3.2 Améliorations du système de post-édition statistique de référence

Ajustement des poids du modèle de post-édition

Comme présenté précédemment, notre système de post-édition statistique repose sur une modélisation log-linéaire composée de 14 modèles. La contribution de chaque modèle du système est estimée par une pondération et l'ensemble des pondérations des modèles constituent les paramètres ou poids du système. Ces poids sont généralement ajustés sur un corpus de développement afin d'être adaptés au corpus et au couple de langues choisis.

Dans nos expérimentations, l'ajustement des poids du système de post-édition statistique est réalisé à l'aide de la méthode MERT sur le corpus de développement de 1 000 énoncés, présenté dans la partie 1. Nous comparerons un ajustement réalisé en

utilisant comme référence les traduction professionnelles *gold-standard* fournies avec le corpus parallèle (DEV_{std}) et celui réalisé en prenant comme référence les hypothèses de traduction de notre système post-éditées manuellement (DEV_{corr}).

Les résultats sont donnés dans le tableau V.6 page 89 (ligne (2)). L'ajustement des poids sur les traductions gold-standard permet d'obtenir, sur le corpus de test, un score BLEU de 60,5 (+ 0,83 % par rapport au système non optimisé) et un score TER de 23,1 (- 4,14 % par rapport au système non optimisé) alors que l'ajustement sur les post-éditions permet d'atteindre un score BLEU de 61,3 (+ 2,16 % par rapport au système non optimisé) et un score TER de 23,4 (- 2,90 % par rapport au système non optimisé). Dans le cas de notre système de post-édition statistique, l'ajustement des poids du modèle à l'aide de la méthode MERT permet bien d'augmenter la qualité des résultats sur le corpus de test et ce, quelque soit le type de référence utilisée (DEV_{corr} ou DEV_{std}). L'ajustement fait sur les données post-éditées permet d'obtenir un gain significativement plus important en terme de score BLEU.

Fonctions de trait	avec DEV_{std}	avec DEV_{corr}	Variation
Nombre d'itérations	9	6	-
Score BLEU	18,02 → 26,39	37,35 → 61,44	-
$T(s/t)$	0.027	0.007	- 0.020
$P_{lex}(s/t)$	0.020	0.018	≈
$T(t/s)$	0.019	0.003	- 0.016
$P_{lex}(t/s)$	0.355	0.034	**- 0.321**
pp	0.049	0.228	**+ 0.178**
wp	-0.049	-0.044	≈
$P(t)$	0.034	0.005	- 0.029
d	0.041	0.186	**+ 0.145**
D_{mp}	0.021	0.010	- 0.011
D_{mf}	0.072	0.162	**+ 0.090**
D_{sp}	0.075	0.013	**- 0.062**
D_{sf}	- 0.107	- 0.006	**+ 0.101**
D_{dp}	0.039	0.171	**+ 0.132**
D_{df}	0.088	0.110	+ 0.022

TABLE V.5 – Différences entre les valeurs de poids issues de l'ajustement sur les traductions professionnelles *gold-standard* (DEV_{std}) *versus* sur les post-éditions (DEV_{corr}) pour un même corpus de développement de 1000 énoncés.

L'analyse des différences entre les valeurs des poids issus de l'optimisation sur les traductions *gold-standard* (DEV_{std}) et ceux issus de l'optimisation sur les post-éditions (DEV_{corr}) montre que l'ajustement fait sur les post-éditions :

- accorde beaucoup moins d'importance à la probabilité de traduction (au niveau des mots) des énoncés post-édités sachant les hypothèses de traduction ($P_{lex}(t/s)$) ;
- pénalise fortement le nombre de segments utilisés pour produire l'hypothèse de traduction. Cela a pour effet de minimiser le nombre de segments utilisés en

cherchant à augmenter leur longueur (*pp*) ;
- autorise moins de distorsion (**d**) ;
- présente des poids liés à la distorsion différents de ceux obtenus par l'ajustement sur les traductions *gold-standard*.

Filtrage de la table de traduction

Par défaut, la table de traduction du modèle de post-édition automatique statistique contient toutes les paires de segments trouvés dans le corpus parallèle. En fonction des données présentes dans le corpus d'apprentissage, cette table peut inclure des données non-informatives, dites « bruitées » (dues, par exemple, à des erreurs d'alignement ou des cas exceptionnels). Pour réduire ce bruit, Johnson et al. suggèrent une méhode consistant à élaguer les paires de segments indésirables à l'origine des données bruitées de la table de traduction [Johnson et al. 2007].

En appliquant ce filtrage sur la table de traduction de notre système de post-édition statistique nous obtenons un gain de 3,5 % de score BLEU et 6,2 % de score TER (par rapport au système de post-édition de référence) avec seulement 6 % des hypothèses de traduction post-éditées par le système (ligne (3) du tableau V.6 page 89).

Interpolation du modèle de langage

Nous tentons maintenant d'enrichir le modèle de langage du système de traduction de référence avec les post-éditions collectées. Le modèle de langage du système de post-édition de référence est identique à celui du corpus de traduction de référence (chapitre II). Il est appris sur un corpus monolingue anglais de 48 653 884 phrases isues des données du Parlement Européen, des Nations Unis et de divers sites Web journalistiques (voir tableau II.1 page 39). Pour cela nous entraînons un modèle de langage sur les post-éditions du corpus d'apprentissage (désigné par LM_{pe}) que nous combinons avec le modèle de langage du système de post-édition de référence (désigné par LM_{ref}). La combinaison des deux modèles de langage est réalisée par interpolation linéaire. Le poids de chacun des deux modèles est attribué « à la main ». Nous testons trois configurations : le modèle de langage appris sur les post-éditions a un poids très important par rapport au modèle de langage de référence ($0.1LM_{ref}$+$0.9LM_{pe}$), l'inverse ($0.9LM_{ref}$+$0.1LM_{pe}$), et les deux modèles de langage ont un poids identique ($0.5LM_{ref}$+$0.5LM_{pe}$). Les résultats associés à chacune de ces configurations sont donnés dans le tableau II.1. Nous remarquons que l'ajout d'un modèle de langage appris sur les post-édition ne permet pas d'améliorer significativement la qualité des traductions en termes de score BLEU et TER et ce, quelque soit le poids accordé à celui-ci ($0,5$, $0,1$ ou $0,9$).

Système de post-édition hiérarchique

Dans cette partie, nous expérimentons une architecture de système de post-édition statistique légèrement différente de celle utilisée jusqu'à maintenant en utilisant une approche basée sur les modèles de segments hiérarchiques (aussi appelée *hierarchical phrase-based models* en anglais) [Chiang 2007]. Ces modèles utilisent une grammaire

hors contexte synchrone (ou SCFG pour *Synchronous Context-Free Grammar* en anglais) qui se compose d'un ensemble de règles où les séquences de mots (symboles terminaux) côtoient des symboles non terminaux pouvant représenter d'autres expressions. Chaque règle de production est représentée par une séquence de mots hiérarchiques et associée à un ensemble de neuf scores de traduction.

Cette approche est connue pour avoir l'avantage de mieux modéliser le réordonnancement et de permettre une meilleure généralisation des observations.

Le système est appris sur les mêmes corpus que ceux utilisés pour le modèle à base de segments. Les poids sont ajustés à l'aide de la technique MERT sur le corpus de développement contenant les post-éditions. Un extrait de la table de règles obtenue est donné dans la figure V.2.

$[X_1]$ draw attention to $[X_2] \rightarrow [X_1]$ draw attention on $[X_2]$
$[X_1]$ of $[X_2]$ involved in $[X_3] \rightarrow [X_1]$ of $[X_2]$ engaged in $[X_3]$
$[X_1]$ reserved seats $[X_2]$ $[X_3] \rightarrow [X_1]$ save the seats $[X_2]$ $[X_3]$
$[X_1]$ am very $[X_2] \rightarrow [X_1]$ am really $[X_2]$
$[X_1]$ aspects $[X_2]$ matter $[X_3] \rightarrow [X_1]$ aspects $[X_2]$ affair $[X_3]$

FIGURE V.2 – Extrait de la table de règles du système de post-édition automatique à base de segments hiérarchiques.

Lors du décodage du corpus de test, 229 énoncés sur 1200 ou 20 % des hypothèses de traduction du système de référence ont été modifiées par le système de post-édition hiérarchique. Les résultats, présentés ligne (5) du tableau V.6 page 89, montrent un score BLEU de 62 et un score TER de 22,7 soit une amélioration des scores de, respectivement, 3,4 % et 5,8 % par rapport au système de post-édition de référence.

Intégration du contexte source

Une brève analyse subjective des résultats obtenus par nos systèmes de post-édition statistique fait ressortir le fait que la post-édition automatique est parfois effectuée « à tort » et à pour conséquence de dégrader la qualité de la traduction voire de créer des contre-sens. Pour tenter de pallier à ce défaut du système, l'idée est ici d'intégrer le contexte source originel de traduction lors de la post-édition statistique.

Pour y parvenir, nous créons un nouveau corpus où les hypothèses de traduction sont mises en correspondance avec la donnée source considérée lors du décodage. Cet alignement est réalisé au niveau des mots d'une part, et des segments d'autre part en accolant l'entité issue de la source et de l'hypothèse avec le signe « # » et les mots d'un même segment avec le signe « _ ». A chaque mot/segment de l'hypothèse fournie par le système de traduction de référence est donc associé le mot/segment source considéré pour sa traduction. Le corpus obtenu (où les données sont de la forme *source#hypothèse*) est alors utilisé comme corpus source pour l'apprentissage du système de post-édition statistique. Les alignements en mots $M_{source}\#M_{hypothese}$ sont obtenus avec l'outil GIZA++ (précédement présenté) et les alignements en seg-

ments $S_{source}\#S_{hypothese}$ avec l'outil d'alignement en segment interne à la boite à outil Moses.

Néanmoins, les alignements des informations de contexte ainsi créés empiriquement ne sont pas toujours fiables et l'apprentissage d'un système de post-édition automatique sur de telles données accroît considérablement la taille du vocabulaire du système ce qui peut potentiellement avoir des effets négatifs sur la qualité de la traduction. Pour tenter d'atténuer ces biais, nous expérimentons donc, en plus de l'intégration systématique du contexte par alignement *source#hypothèse*, un seuillage de ces information en fonction de la vraisemblance de leur alignement donnée par l'estimation v (fournie par les outils qui créent les alignements). Par exemple, un seuil de $v \geq 0,8$ pour les alignements en mots consiste à ne considérer que les alignements $M_{source}\#M_{hypothese}$ dont la probabilité donnée par GIZA++ est supérieure ou égale à $0,8$.

L'intégration du contexte source au niveau mot et phrase, avec et sans seuillage est réalisé sur l'ensemble des énoncés des corpus d'apprentissage, de développement et de test. Le traitement d'une phrase est donné comme exemple dans la figure V.3.

Après évaluation, le meilleur résultat des systèmes de post-édition incluant le contexte source est obtenu avec celui appris avec l'information contextuelle au niveau mot et un seuil sur la probabilité d'alignement à 0,8. Les résultats de l'évaluation (score BLEU de 17,7 et score TER de 42 indiqués ligne (6) du tableau V.6 page 89) indique une qualité de traduction bien au deça de celle obtenue avec les précédents systèmes de post-édition. Ceci peut être expliqué en partie par l'importante hétérogénéité obtenue par de telles données, en partie au niveau de la table de traduction.

Phrase source : Dans un monde idéal, les conflits se règlent à force d' accords et de traités.

Hypothèse de traduction : In an ideal world conflicts are resolved to force of agreements and treaties.

$M_{source}\#M_{hypothese}$: Dans#in un#a monde#world idéal#ideal ,#, les_conflits#conflicts se_règlent#settle à#to force#force d'#of accords#agreements et#and traités#treaties .#.

$S_{source}\#S_{hypothese}$: Dans_un_monde#in_a_world idéal_,#ideal_, les_conflits#conflicts se_règlent#settle à_force#to_force d'_accords_et_de#of_agreements_and traités_.#treaties_.

Seuil à 0,8 sur $M_{source}\#M_{hypothese}$: Dans#in un#a monde#world idéal#ideal ,#, les_conflits#conflicts settle to force#force of accords#agreements et#and treaties .#.

Seuil à -2 sur $S_{source}\#S_{hypothese}$: Dans_un_monde#in_a_world idéal_,#ideal_, conflicts settle to force of agreements and traités_.#treaties_.

FIGURE V.3 – Exemple de considération de l'information source dans le corpus utilisé pour la post-édition automatique statistique.

Spécificité du système de post-édition automatique	Ajustement des poids	Modèle de langage	Filtrage de la TT	Score BLEU	Score TER	Enoncés modifiés par le SPES
(1) Système de référence	/	LM_{ref}	/	60.0	24,1	36 %
(2) Ajustement des poids	DEV_{std}	LM_{ref}	/	60.5	23.1	32 %
	DEV_{corr}	LM_{ref}	/	61.3	23.4	22 %
(3) Filtrage de la TT [1]	DEV_{corr}	LM_{ref}	Oui	62.1	22.6	6 %
(4) Interpolation des modèles de langage	DEV_{corr}	$0.5LM_{ref}$+$0.5LM_{pe}$	/	60.1	24.0	38 %
	DEV_{corr}	$0.9LM_{ref}$+$0.1LM_{pe}$	/	60.1	23.9	31 %
	DEV_{corr}	$0.1LM_{ref}$+$0.9LM_{pe}$	/	59.7	24.2	48 %
(5) Système hiérarchique	DEV_{corr}	LM_{ref}	/	62.0	22.7	20 %
(6) Avec contexte source [2]	DEV_{corr}	LM_{ref}	/	42.0	17.7	30 %

TABLE V.6 – Evaluations des différents systèmes de post-édition automatique statistique (appliqués sur notre système de traduction de référence présenté dans la partie II).

3.3 Sélection des phrases à post-éditer automatiquement

L'idée est ici d'améliorer les résultats obtenus avec le système de post-édition automatique en évitant de dégrader la qualité des hypothèses de traduction lors de la post-édition automatique.

Pour cela, nous proposons de sélectionner les hypothèses de traduction du système de référence susceptibles d'être améliorées par une post-édition automatique. L'application du système de post-édition statistique, systématique dans les expérimentations précédentes, serait alors sélective et ciblée.

Dans leurs travaux, [Suzuki 2011] et [Rubino et al. 2012] proposent d'utiliser un classifieur pour déterminer les phrases à post-éditer automatiquement. Les expériences menées sur le SPES appliqué de façon sélective permettent d'améliorer la performance par rapport à l'utilisation systématique de la post-édition.

Evaluation du potentiel

Avant toute chose, nous souhaitons évaluer le potentiel d'une sélection, au niveau des phrases, des hypothèses de traduction à post-éditer automatiquement. Nous proposons d'évaluer le potentiel de la post-édition automatique sélective avec une méthode dite « oracle » : nous évaluons, dans un premier temps, les hypothèses de traduction du corpus

1. TT = Table de Traduction.
2. Au niveau mot et avec un seuil sur les probabilités d'alignement à 0,8.

de test avant et après post-édition puis nous effectuons, dans un deuxième temps, une sélection phrase-à-phrase afin d'obtenir un score maximal.

L'évaluation au niveau des phrases est effectuée à l'aide du score BLEU modifié appliqué sur les hypothèses de traduction du système de référence d'une part ($BLEU_{hyp}$) et sur leurs post-éditions automatiques d'autre part ($BLEU_{spe}$). Nous calculons ensuite la différence entre ces deux scores BLEU selon l'équation V.1 afin de juger si la post-édition automatique améliore ou dégrade le score BLEU de l'hypothèse de traduction.

$$\Delta BLEU = BLEU_{spe} - BLEU_{hyp} \tag{V.1}$$

Nous distinguons alors trois cas :
- $\Delta BLEU > 0$: si la post-édition améliore la qualité de l'hypothèse de traduction (en termes de score BLEU) ;
- $\Delta BLEU < 0$: si la post-édition détériore la qualité de l'hypothèse de traduction (en termes de score BLEU) ;
- $\Delta BLEU = 0$: si la post-édition ne modifie pas la qualité de l'hypothèse de traduction (en termes de score BLEU).

Nous calculons ensuite la performance sur le corpus de test en ne considérant que les phrases post-édités pour lesquelles la post-édition améliore la qualité de la phrase en termes de score BLEU.

$$BLEU_{sentence} = \begin{cases} BLEU_{spe} & \text{si} \quad \Delta BLEU \geq 0 \\ BLEU_{hyp} & \text{sinon} \end{cases} \tag{V.2}$$

Résultats

Le résultat de la post-édition statistique sélective est donné dans le tableau V.7. A noter qu'il s'agit d'un score « oracle », c'est à dire que la sélection des phrases automatiquement post-éditées est faite *a posteriori*. Ce score évalue le potentiel d'une méthode de post-édition sélective.

Le système de post-édition utilisé ici est le système de référence dont les poids ont été ajustés sur le corpus post-édité (ligne (2) du tableau V.6).

Même si la post-édition sélective faite au niveau des phrases (qui obtient un score BLEU de 62,16) est significativement meilleure que la post-édition systématique (qui obtient un score BLEU de 61,33) celle-ci ne permet pas d'outrepasser significativement la performance du système du système de traduction de référence (qui obtient un score BLEU de 62,10).

L'origine des traductions sélectionnées pour obtenir le score oracle de la post-édition sélective est donnée dans le tableau V.8.

Les résultats montrent que le gain à espérer en sélectionnant les phrases à post-éditer sur la base du score BLEU est faible. Cependant, les scores obtenus nous permettent d'observer que la post-édition sélective des hypothèses de traduction obtient des résultats de meilleurs qualité que la post-édition systématique des sorties de traduction.

	Score BLEU
PBMT	62,10
Post-édition systématique	61,33
Post-édition sélective (oracle)	62,16

TABLE V.7 – Evaluation de la qualité des traduction selon la méthode de post-édition appliquée au sorties du PBMT de référence (sur le corpus de test de 1200 phrases).

Scores	Proportion de phrases
$\Delta BLEU > 0$	1,9 % (23/1200)
$\Delta BLEU < 0$	13,7 % (165/1200)
$\Delta BLEU = 0$	84,3 % (1012/1200)

TABLE V.8 – Distribution des énoncés du corpus de test (1200 phrases) selon la préférence estimée en terme de score BLEU entre l'hypothèse de traduction du système ($BLEU_{hyp}$) d'une part et sa post-édition automatique d'autre part ($BLEU_{spe}$)

4 Conclusion

L'idée était ici d'utiliser les phrases corrigées (ou post-éditions) précédemment collectées comme une source d'information additionnelle à ré-intégrer dans le système de traduction pour produire d'autres hypothèses que l'on espère améliorées.

Après avoir partitionné le corpus collectée en trois sous-ensembles (pour l'apprentissage, le développement et le test des systèmes), nous avons évalué deux protocoles expérimentaux : l'enrichissement du système de traduction de référence par ajout d'une table de traduction apprise sur les post-éditions et la post-édition automatique des résultats du système de référence.

La première approche, qui consiste à ré-intégrer les post-éditions en enrichissant le modèle de traduction du système de référence avec une table de traduction apprise sur les post-éditions collectées, ne montre pas d'amélioration des scores automatiques par rapport au système de référence. Les résultats de la deuxième approche, par post-édition automatique des résultats de traduction du système de référence, montrent que celle-ci constitue une piste de travail intéressante pour ré-intégrer les post-éditions humaines dans le processus de traduction.

Même si l'usage d'un post-éditeur automatique « naïf » dégrade significativement la qualité des sorties « brutes » du système de traduction de référence, nous avons expérimenté plusieurs pistes d'adaptation du système afin de mieux évaluer le potentiel de l'approche.

Au delà de l'application systématique de la post-édition statistique, nous expérimentons la post-édition sélective des hypothèses de traduction. Le score « oracle » (sélection des phrases *a posteriori*) calculé pour évaluer le potentiel de la méthode de post-édition sélective se montre significativement meilleur que celui obtenu avec la post-édition systématique, mais les résultats obtenus montrent cependant que le gain à

espérer en développant une méthode de sélection automatique des phrases à post-éditer reste faible.

Chapitre VI

Etude approfondie de systèmes de post-edition automatiques probabilistes

A la lumière des performances peu satisfaisantes des premières expérimentation présentées dans le chapître V, nous proposons, par la suite, d'étudier et de fournir une meilleure compréhension des systèmes de post-éditions statistiques quand ils sont utilisés pour améliorer les sorties de système de traduction probabilistes.

Après une étude bibliographique sur les travaux relatifs aux systèmes de post-edition automatiques (section 1), nous comparerons la qualité des résultats du système de post-édition statistique lorsqu'il est entraîné sur un corpus de post-éditions humaines (configuration « réelle » dans notre scénario de ré-apprentissage) et lorsqu'il est appris sur un corpus de post-éditions simulées (configuration « simulée ») par des traductions professionnelles indépendantes (section 2). Nous nous intéresserons enfin à l'utilisation du système de post-édition statistique pour améliorer un système de traduction probabiliste générique et comparerons les résultats obtenus avec ceux produits par un système de post-édition statistique spécialisé utilisé pour adapter un système à un domaine donné (section 3).

1 Etude bibliographique de la post-édition automatique

1.1 Traduction professionnelle et post-édition

La tâche de post-édition consiste à éditer la sortie textuelle d'un système automatique dans le but de la corriger. Dans le domaine de la traduction automatique, la post-édition manuelle des résultats du système est utilisée depuis des années et consiste en un post-traitement à la traduction automatique qui a pour but de corriger les traductions faites par le système de traduction pour produire des traductions de meilleure qualité.

De nombreuses études ont montré l'efficacité de l'utilisation d'un système de traduction automatique combiné à une post-édition manuelle, dans un but de diffusion de données textuelles. Des expérimentations menées par [Garcia 2011] ont montré que même si la post-édition de sorties « brutes » de systèmes de traduction automatique n'apporte pas nécessairement de gain en termes de productivité, cela aide à produire des traductions significativement meilleures comparées à une traduction qui serait faite directement à partir du texte source (ceci, quelque soit les langues et la direction de traduction, la difficulté du texte ou l'expérience du traducteur). Plus récemment, l'entreprise Autodesk a réalisé une expérience à grande échelle pour tester l'influence de l'usage de systèmes de traduction automatique sur la productivité des traducteurs professionnels. Les résultats [41] montrent que la post-édition de résultats provenant de systèmes automatiques améliore significativement la productivité des traducteurs comparativement aux traductions réalisées sans l'aide de systèmes automatiques. Ces résultats ne sont influencés ni par la paire de langues, ni par l'expérience ou la préférence du traducteur (post-édition ou traduction à partir de la phrase source), ni par la taille de la phrase à traduire.

1.2 Automatisation de la tâche de post-édition

La tâche de post-édition est à l'origine manuelle : la traduction à corriger est éditée via une interface d'édition puis corrigée par un annotateur humain.

Ce processus de post-édition manuelle est coûteux à mettre en œuvre, il requiert des annotateurs experts de la langue cible, des moyens logiciels et logistiques conséquents et le temps d'annotation reste relativement long. Pour faciliter cette tâche, est apparue l'idée de l'automatiser, c'est-à-dire d'apprendre le comportement des annotateurs humains pour pouvoir propager automatiquement les corrections, effectuées par des professionnels, sur de nouvelles traductions. Il convient de noter que cette post-édition automatique n'a pas pour objectif de remplacer la post-édition humaine mais de la simplifier et/ou d'en augmenter la productivité.

C'est dans ce but qu'au début des années 90, [Knight et Chander 1994] ont proposé un système automatique reposant sur la post-édition afin d'aider à la sélection d'articles lors de la traduction de textes du japonais vers l'anglais. Plus tard, [Allen et Hogan 2000] développèrent un module automatique de post-édition à base de règles capable de capturer et corriger *« les erreurs fréquentes et répétées produites par les systèmes de traduction automatique à base de règles »*.

[Elming 2006] fut le premier à proposer et évaluer un module de post-édition automatique. Dans ses expérimentations, J. Elming fait traduire des textes représentant des brevets de chimie (domaine spécialisé) par un système de traduction à base de règles nommé *Patrans* dont les traductions brutes sont ensuite corrigées avec un module de post-édition automatique (« transformation-based ») appris sur 12 000 traductions post-éditées manuellement. Il montre une amélioration significative de la qualité de la traduction avec l'utilisation de son module de post-édition automatique. De la même façon [Guzmàn 2007] propose d'identifier des « patrons » (ou *patterns*) d'erreurs lin-

41. http ://translate.autodesk.com/productivity.html

guistiques propres à un couple de langues donné (ici anglais/espagnol) puis d'utiliser des expressions régulières pour corriger les sorties d'un système de traduction automatique à base de règles.

Parallèlement à cela, la disponibilité croissante de sorties brutes de systèmes de traduction automatiques alignées avec des post-éditions manuelles de traducteurs professionnels a fait émerger l'idée de post-édition statistique.

1.3 Post-édition statistique

La tâche de post-édition peut être vue comme une tâche de traduction entre une sortie brute d'un système et cette sortie « corrigée ». Les systèmes de post-édition statistique (SPES) sont alors appris comme des systèmes de traduction automatique monolingues pour lesquels les hypothèses de traduction jouent le rôle de la langue source et les post-éditions humaines celui de la langue cible. Ce sont des systèmes, appris empiriquement sur de grands corpus de données, ayant pour but de modéliser et d'automatiser la tâche de post-édition.

En 2007, [Simard et al. 2007a] sont les premiers à proposer l'utilisation d'un système de traduction statistique basé sur les segments (PBMT) dans un objectif de post-édition statistique. Dans ce cadre, le système a pour but d'apprendre des « règles de correction » entre les hypothèses de traduction issues d'un système de traduction automatique (langage source du SPES) et leurs versions corrigées (langage cible du SPES). Une telle approche rend alors les SPES facile à apprendre et à optimiser sur de nouvelles données d'apprentissage.

[Simard et al. 2007a] ont montré avec succès l'efficacité d'un SPES (appris avec le PBMT *Portage*) pour améliorer les sorties d'un système de traduction commercial à base de règles. Les expérimentations ont été menées sur des textes issus d'un domaine spécifique (un site Web canadien d'offre d'emplois [42]) et le SPES est appris sur 35 000 traductions manuellement post-éditées. Encouragé par ces résultats, Simard tenta de corriger, de la même façon, les sorties d'un système PBMT (appris à l'aide de *Portage*) à l'aide d'un système de post-édition statistique mais ne constata aucune amélioration de la traduction avec cette configuration.

De la même manière, les études décrites dans [Isabelle et al. 2007], [Simard et al. 2007b] et [Diaz de Ilarraza et al. 2008] ont montré qu'un système de traduction à base de règles cascadé avec un système de post-édition statistique produisaient de bien meilleurs résultats que ceux obtenus par les systèmes de traduction seuls.

Les études sur les systèmes de post-édition statistiques se sont majoritairement focalisées sur des architectures combinant SPES et système de traduction à base de règles. Peu d'études (parmi celles-ci peuvent être citées [Béchara et al. 2011, Diaz de Ilarraza et al. 2008, Lagarda et al. 2009]), se sont intéressées à étudier l'utilisation de SPES pour améliorer la sortie de systèmes de traduction statistiques à base de segments.

42. http ://www.jobbank.gc.ca

2 Corpus post-édité réel *vs* simulé pour l'apprentissage du système

Dans un scénario applicatif d'amélioration incrémentale, les SPES utilisent comme traduction cible les post-éditions manuelles des hypothèses de traduction du système. Lorsque ces post-éditions manuelles sont remplacées par des traductions pré-existantes (souvent produites, indépendamment de toute post-édition, par des traducteurs professionnels), nous parlerons de « post-édition simulée » en contraste avec la « post-édition réelle » lorsque les traductions cibles sont des hypothèses de traduction manuellement post-éditées. Il est important de noter que la « post-édition réelle » correspond au scénario implémenté dans les situations réelles (quand les rétro-actions des utilisateurs sont ré-utilisées pour améliorer un système donné) alors que la « post-édition simulée » donne accès à beaucoup plus de données d'apprentissage (utilisation de corpus pré-traduits).

2.1 Travaux antérieurs

Dans la grande majorité des études, les conditions expérimentales mettent en oeuvre un apprentissage du système de post-édition statistique avec des post-éditions simulées. En effet, il est fréquent d'utiliser un corpus bilingue pré-traduit pour simuler l'annotation et ne pas avoir à créer ou utiliser d'outil de post-édition : les phrases cibles du corpus parallèle sont alors utilisées pour simuler la production d'un post-éditeur humain.

Comme dans [Dugast et al. 2007, Lagarda et al. 2009, Dugast et al. 2009, Béchara et al. 2011], plusieurs travaux ont tenté de montrer que les SPES peuvent être appris avec succès sur des traductions humaines pré-existantes plutôt que des traductions post-éditées spécifiques au système.

[Kuhn et al. 2010] expérimentent les deux configurations : apprentissage sur données réelles (les hypothèses de traduction du système alignées avec leurs post-éditions humaines) et apprentissage sur données simulées (les hypothèses de traduction du système alignées avec les traductions de référence du corpus). Les deux systèmes (SPES « réel » et SPES « simulé ») obtiennent de bons résultats mais les performances respectives ne sont pas comparables car ni le système de traduction de référence ni le corpus d'apprentissage du SPES (en termes de taille et domaine) ne sont identiques dans les deux cas.

A notre connaissance, aucun travail ne compare les deux approches (SPES « réel » *vs* SPES « simulé ») en utilisant un même corpus de données source pour apprendre les deux systèmes de post-édition statistique. En considérant un même ensemble de données en langue source, est-ce qu'un SPES appris sur un corpus de post-édition « simulée » est aussi efficace qu'un SPES appris sur un corpus de post-édition « réelle » ? C'est ce à quoi nous allons tenter de répondre dans les expérimentations qui vont suivre.

2.2 Expérimentations

Afin de créer deux systèmes de post-édition automatique comparables, nous utilisons un corpus source d'apprentissage identique du côté des données source pour les deux cas (correspondant à celui décrit dans la partie 3.1) et aligné du côté cible avec les hypothèses de traduction de notre système post-éditées pour l'un (configuration « réelle ») et les traductions de référence du corpus parallèle pour l'autre (configuration « simulée »).

Les deux SPES ainsi obtenus sont ensuite appliqués sur un même ensemble de sorties du système de traduction de référence et leurs performances sont évaluées sur le corpus de test (1 200 phrases) en utilisant la même distinction que celle utilisée lors de l'apprentissage : nous utilisons comme référence les post-éditions du corpus de test pour la configuration « réelle » et les traductions professionnelles pour la configuration « simulée ». Il est à noter que le modèle de langage cible est identique dans les deux cas. L'hypothèse initiale est que l'utilisation du corpus de post-édition devrait permettre d'aboutir à de meilleurs résultats que l'utilisation du corpus de traductions professionnelles indépendantes en raison de la « proximité » entre les hypothèses de traductions brutes et les traductions post-éditées correspondantes.

2.3 Résultats

Comme présenté dans le tableau VI.1, les sorties brutes du système de traduction de référence (nommé « PBMT » ici) obtiennent un score TER de 22,8 quand elles sont comparées avec les post-éditions humaines et 55,3 quand elles sont comparées avec les traductions de référence indépendantes. Un score TER de 22,8 signifie qu'environ 22,8 % des mots (ou segments) de l'hypothèse de traduction ont dû être édités (supprimés, déplacés ou ajoutés) pour produire la traduction de référence.

La différence entre les scores de la configuration « simulée » et ceux de la configuration « réelle » est due au fait que les références utilisées pour l'évaluation sont différentes : il s'agit respectivement, dans le premier cas de traductions professionnelles indépendantes et, dans le deuxième cas, de post-éditions humaines plus proches des hypothèses de traductions du système.

La post-édition statistique avec le système appris sur les données « réelles » et appliquée sur les sorties de notre système de traduction de référence, conduit à une légère dégradation des scores TER (de 22,8 pour les sorties brutes à 23,4 après post-édition statistique) et BLEU (de 62,1 pour les sorties brutes à 61,3 après post-édition statistique). Néanmoins, selon [Koehn 2004], ces différences ne sont pas suffisantes pour être significatives.

Alors que le SPES appris sur les données « réelles » ne dégrade pas significativement les résultats de traduction, l'application du SPES appris sur les données « simulées », d'un autre côté, entraîne une dégradation significative des résultats : après post-édition statistique, la qualité de la traduction perd, sur le corpus test d'évaluation, 4 % de son score TER et 6 % de son score BLEU.

Compte tenu de nos paramètres expérimentaux (*i.e.* un corpus d'un domaine général et de taille moyenne) nos résultats ne montrent aucune amélioration de la qualité de la

Système	Configuration « simulée »	Configuration « réelle »
PBMT	55.3 (*26.5*)	22.8 (*62.1*)
PBMT + SPES	57.5 (*25.0*)	23.4 (*61.3*)

TABLE VI.1 – Performance — scores TER (*BLEU*) — selon l'utilisation de données de post-éditions « simulées » *vs* « réelles » pour l'apprentissage du SPES

traduction après post-édition automatique et ce, quelque soit la configuration choisie (post-édition « réelle » ou « simulée »).

2.4 Apprentissage à grande échelle

Pour compléter notre étude précédente, nous avons étudié l'impact de la taille du corpus d'apprentissage sur la performance du système de post-édition statistique. Compte tenu de la taille modérée du corpus de post-éditions humaines disponible (10 881 phrases), nous avons utilisé un corpus de plus grande échelle avec la configuration « simulée » pour réaliser des expériences à grande échelle.

Nous employons le corpus bilingue français/anglais des Nations Unies qui est composé de textes des résolutions prises par l'Assemblée générale des Nations Unies, traduits par des professionnels. Dans le milieu de la traduction automatique probabiliste, ce corpus est traditionnellement utilisé comme un corpus d'apprentissage de grande taille pour le domaine général [43].

Nous avons aussi considéré le corpus journalistique de 8 681 phrases (10k) (voir la partie 3.1) et divisé le corpus des Nations Unies pour obtenir un corpus de 50 000 phrases (50k), 100 000 phrases (100k), 500 000 phrases (500k), 1 000 000 phrases (1M) et 2 000 000 phrases (2M). Chacun des corpus mentionné ici inclut le corpus journalistique 10k. Nous avons ensuite utilisé ces six corpus pour entraîner six systèmes de post-édition probabiliste. Le seul paramètre qui différencie les six systèmes produits est la taille du corpus d'apprentissage (le modèle de langage étant le même).

Nous avons évalué les six systèmes de post-édition sur l'ensemble de test et reporté les performances, en termes de scores BLEU et TER, sur la figure VI.1 (les systèmes sont classés en fonction de la taille de leur corpus d'apprentissage). Les résultats ne montrent pas de gains significatifs, ni pour le score TER, ni pour le score BLEU, avec l'augmentation de la taille du corpus d'apprentissage. En d'autres termes, dans un contexte de traduction français/anglais de domaine général, l'ajout de données d'apprentissage supplémentaires n'améliore pas les résultat du système de post-édition.

43. Le corpus est disponible à l'URL : http ://www.statmt.org/wmt12/translation-task.html

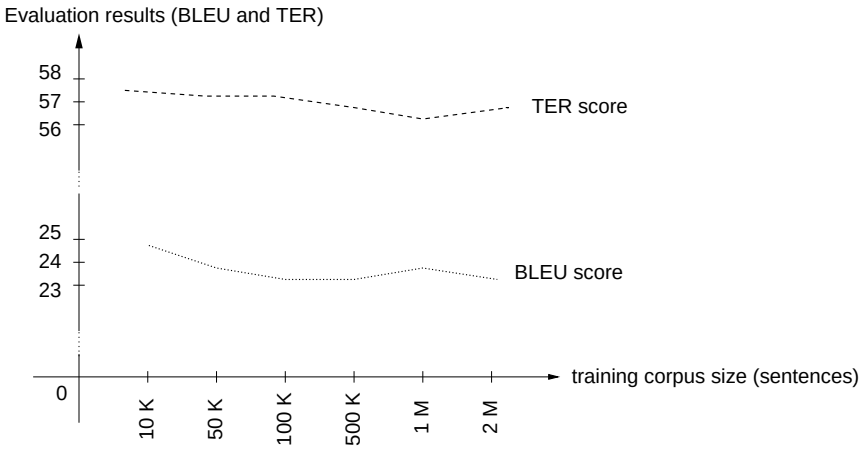


FIGURE VI.1 – Performances — scores TER et BLEU — de SPES appris sur une configuration « simulée » en fonction de la taille du corpus d'apprentissage (en phrases)

3 Post-édition en domaine général *vs* spécialisé

3.1 Travaux antérieurs

Alors que les systèmes de post-édition statistique n'ont pas réussi à montrer leur efficacité pour améliorer significativement les résultats d'un système de traduction probabiliste générique, des études sur les SPES ont montré leurs intérêts comme méthode d'adaptation au domaine. Dans leurs travaux, [Isabelle et al. 2007] et [Simard et al. 2007b] ont montré que les SPES appris sur des données spécifiques à un domaine pouvaient être utilisés pour adapter un système de traduction probabiliste général à un nouveau domaine de spécialité.

Lors de leurs expériences, [Diaz de Ilarraza et al. 2008] ont remarqué que si l'application d'un SPES aux résultats d'un système de traduction automatique à base de règles (RBMT) est efficace pour l'adaptation au domaine, celui-ci se révèle inefficace s'il est appliqué sur les sorties d'un système de traduction empirique (PBMT). Dans leurs travaux, [Lagarda et al. 2009] et [Béchara et al. 2011] aboutissent à la même conclusion lorsqu'ils appliquent un SPES spécifique à un domaine sur les sorties d'un système PBMT de domaine général. [Béchara et al. 2011], quant à eux, proposent quelques pistes afin de spécialiser le SPES afin d'améliorer ses résultats sur un PBMT dans une tâche d'adaptation au domaine (ils proposent, par exemple, de tenir compte des informations de contexte issues de la source lors de la post-édition).

Même si ces études confirment l'efficacité des SPES lorsqu'ils sont appliqués aux sorties de RBMT à des fins d'adaptation au domaine, celles-ci montrent peu de résultats positifs de SPES appliqués aux systèmes PBMT. Comme nous l'avons montré précédemment dans notre étude, un SPES utilisé dans un contexte de domaine général n'apporte aucune amélioration lorsqu'il est appliqué sur les résultats d'un système PBMT. Si les SPES ne peuvent corriger efficacement les systèmes PBMT, peuvent-ils être utilisés pour adapter ces mêmes systèmes à un nouveau domaine ? Pour répondre à

cette question, nous avons mis au point un protocole expérimental pour tester le potentiel d'une approche d'adaptation au domaine avec SPES par rapport à une utilisation de ces mêmes systèmes sur un domaine général.

3.2 Expérimentations

Étant donné la nature des données disponibles, les expériences suivantes se dérouleront dans le cadre d'une configuration de post-édition « simulée ». Nous avons utilisé le corpus post-édité décrit dans le chapitre III avec les traductions de référence indépendantes faites par des professionnels et un corpus spécifique au domaine des sciences de l'eau traduit par des professionnels qui ont post-édité des résultats de traducteurs automatiques.

Ce corpus spécialisé est extrait de l'encyclopédie EOLSS (pour *Encyclopedia Of Life Support Systems* en anglais) qui est une base de connaissance de plus de 200 thèmes sur le développement durable, développée avec le soutien de l'UNESCO (officiellement publiée sur Internet [44] et en langue anglaise par le Directeur général de l'UNESCO le 3 septembre 2002). L'extrait utilisé dans nos expérimentations comporte des textes sur la thématique des sciences de l'eau, traduits en Français dans le contexte du projet SECTra_w [Huynh et al. 2008, Blanchon et al. 2009]. Des exemples de phrases issues du corpus spécialisé sont données dans le tableau VI.2.

Phrase source	Traduction de référence
• Cela explique le refroidissement superficiel sous l'évaporation de l'eau.	• This explains surface cooling under water evaporation.
• Ce paradoxe a été expliqué par l'amplification du tsunami à cause d'un glissement de terrain sous-marin juste après le tremblement de terre.	• This paradox was explained by amplification of the tsunami due to a submarine landslide just after the earthquake.
• Le cadmium et le zinc s'accumulent dans les organes reproductifs des plantes.	• Cadmium and zinc are accumulated in generative plant organs.
• L'eau avec une conductivité électrolytique élevée est susceptible de causer un tel affaiblissement.	• Water with high electrolytic conductivity is likely to cause such impairment.

TABLE VI.2 – Exemples de phrases issues du corpus spécialisé (EOLSS)

Les corpus spécialisés et généraux sont mis en parallèle dans le tableau VI.3. Ils sont comparables en termes de taille et ne diffèrent que par la spécificité de leur vocabulaire et de leur grammaire liés à leur domaine d'usage. Comme le corpus de domaine général, le corpus de domaine spécialisé a été divisé en un ensemble d'apprentissage ($\approx$ 9 000 phrases), un ensemble de développement (1 000 phrases), et un ensemble de test (1 200 phrases). Un nouveau système de post-édition statistique a été appris sur les données

44. www.eolss.net

au domaine spécialisé (celui précédemment présenté utilisant des données du domaine général).

Corpus	Spécialisé	Général
Domaine	Sciences de l'eau	Journalistique
Nature	Encyclopédie EOLSS	Divers sites Web
Taille du vocabulaire	14 015 mots	21 982 mots
Taille des phrases	≈ 22 mots	≈ 28 mots
Source	Corpus traduit avec l'outil SECTra_w [Huynh et al. 2008, Blanchon et al. 2009]	Corpus fourni par la campagne d'évaluation internationale WMT [45]

TABLE VI.3 – Comparaison entre le corpus du domaine général *vs* spécialisé

3.3 Résultats

Comme présenté dans le tableau VI.4, le système de traduction automatique de référence appris sur un corpus d'apprentissage du domaine général obtient un score TER de 55,3 sur le corpus de test du domaine général et un score de 46,7 sur le corpus de test du domaine spécialisé. Cela signifie que le système de traduction de référence, bien qu'appris sur des données du domaine général, produit des traductions de meilleure qualité quand il est appliqué sur le corpus de test spécialisé. D'autre part, bien que le SPES appris sur un domaine général n'apporte pas de gain sur le corpus de test de la même nature, le SPES appris sur les données spécialisées améliore de manière significative les sorties du système de traduction de référence sur le corpus de test spécifique au domaine de l'eau : le score TER décroît de 46,7 à 39,2 (-19,2 %) et le score BLEU suit le même tendance en augmentant de 33,3 à 40,1 (+20,6 %).

La première ligne du tableau VI.5 indique que le système de post-édition statistique spécifique au domaine produit non seulement des résultats de meilleure qualité (comme on le voit dans le tableau VI.4) mais il modifie plus de phrases (91 %) que le SPES de domaine général (qui modifie 75 % des phrases). La deuxième ligne du tableau VI.5 indique la proportion des traductions issues du système de traduction de référence améliorées (en terme de score TER) grâce à la post-édition statistique : le SPES appris et appliqué sur les données spécifiques améliore 58 % des sorties du système de référence contre seulement 11 % pour le SPES appris et appliqué sur les données générales. Quelques exemples de traductions des données spécifiques au domaine de l'eau sont présentées, avant et après post-édition, dans le tableau VI.6.

3.4 Adaptation au domaine

Les principales questions soulevées par ces nouvelles expériences sont les suivantes : Pourquoi le SPES améliore avec succès les données spécifiques au domaine et échoue

45. http ://www.statmt.org/wmt10

Système	Domaine spécifique	Domaine général
PBMT	46,7 (*33,3*)	55,3 (*26,5*)
PBMT+$SPES_{specifique}$	39,2 (*40,1*)	/
PBMT+$SPES_{general}$	/	57,5 (*25,0*)

TABLE VI.4 – Performances des SPES — Scores TER (*BLEU*) — selon les domaines d'application

Taux de post-édition	Domaine spécifique	Domaine général
Phrases corrigées	91 %	75 %
Traductions améliorées	58 %	11 %

TABLE VI.5 – Taux de phrases corrigées et de traductions dont la qualité a été améliorée en fonction du domaine d'application

sur des données de nature générale ? Est-ce que les systèmes de post-édition statistiques sont voués à des tâches d'adaptation au domaine ?

Adaptation au domaine par correction lexicale

Dans [Dugast et al. 2007], les résultats d'une classification manuelle des modifications effectuées par le SPES sur la sortie brute du système de traduction permettent de constater que les SPES apportent des améliorations significatives en termes de choix lexical, mais aucune amélioration dans le réordonnancement des mots ou la grammaticalité de la phrase. A partir de ce constat, il semble nécessaire de vérifier quelle est la part des corrections lexicales dans le succès des systèmes de post-éditions probabilistes dans les tâches d'adaptation au domaine.

Par la suite, nous allons procéder à une analyse de la gestion des mots hors-vocabulaire lors de la phase de post-édition statistique. Un mot est considéré « hors-vocabulaire » pour un système de traduction automatique lorsqu'il n'apparait pas dans le dictionnaire de traduction dudit système, dans le cas des systèmes de traduction probabilistes à base de segments, il s'agit de la table de traduction.

Dans nos expériences, nous avons comparé le nombre de mots hors-vocabulaire avant et après post-édition statistique appliquée au domaine général d'une part et spécialisé d'autre part. Les statistiques présentées sont calculées sur un ensemble de 2200 phrases (qui correspond à la concaténation des corpus de développement et de test précédemment utilisés) à la fois pour le domaine général et spécifique.

Les résultats, présentés dans le tableau VI.7, révèlent une proportion équivalente de mots hors-vocabulaire dans les deux ensembles (2,8 % pour le corpus spécifique au domaine et 2,7 % pour le corpus général). Il est toutefois important de noter que les données spécifiques au domaine présentent un ratio type-token [46] de mots hors-vocabulaire de 61 % contre 72 % pour le domaine général. Les données spécifiques

46. Le ratio type/token est une mesure de variabilité du vocabulaire d'un texte. Plus le ratio type/-token est important, plus la variabilité lexicale est grande.

Phrase source	Hypothèse de traduction	Post-édition statistique
• Unité africaine de recherche sur les questions de l'eau	• African **unit of research on issues of water**	• African **water issues research unit**
• Réduction de la salinité des eaux souterraines dans les zones agricoles	• Reducing **the salt content of groundwater** in agricultural areas	• Reducing **groundwater salinity** in agricultural areas
• L'offre est en grande partie déterminée par la productivité dans les zones irriguées et pluviales[...]	• **The offer** is largely determined by productivity in the irrigated **areas and *pluviales***[...]	• **Supply** is largely determined by productivity in the irrigated **and rain-fed areas**[...]
• Atténuation des tsunamis et étapes pour réduire des pertes potentielles	• **Mitigation tsunamis** and steps to reduce **the** potential losses	• **Tsunami mitigation** and steps to reduce potential losses
• Equipement pour le traitement de l'eau	• Equipment for **the treatment of water**	• Equipment for **water treatment**
• Plusieurs exigences sont établies pour l'eau potable	• Several requirements are **set** for **drinking** water	• Several requirements are **established** for **potable** water

TABLE VI.6 – Exemples de traductions issues du domaine spécifique

au domaine de l'eau contiennent donc moins de variation lexicale que celles issues du domaine général.

Enfin, l'application du système de post-édition statistique du domaine spécialisé corrige 56 % des mots hors-vocabulaire contenus dans les résultats du système de traduction de référence contre 7 % pour le SPES général.

Statistiques des mots hors-vocabulaire (H-V)	Domaine spécialisé	Domaine général
Hypothèses de traduction contenant des mots H-V	40 %	43 %
Taux de mots H-V du système	2.8 %	2.7 %
Ratio type-token des mots H-V	61 %	72 %
Mots H-V corrigés par le SPES	56 %	7 %
Noms commums H-V corrigés par le SPES	**42 %**	**1 %**

TABLE VI.7 – Statistiques des mots hors-vocabulaire (H-V) selon le domaine d'application

Afin de mieux comprendre ces résultats, nous avons analysé la nature des mots hors-vocabulaire pour les deux ensembles de données. Les résultats sont présentés dans le tableau VI.8. Les mots hors-vocabulaire du système de traduction de référence sont en grande majorité : des noms communs (75,6 % des mots hors-vocabulaire) pour les données spécifiques et des noms propres ainsi que des mots en langue étrangère (81,5 %

Nature des mots hors-vocabulaire corrigés	Domaine spécialisé	Domaine général
Noms propres	16,8 %	**46,8 %**
Mots en langue étrangère	2,3 %	**34,7 %**
Erreurs dans la phrase source	1,5 %	2,4 %
Nombres	3,3 %	5,6 %
Noms communs	**75,6 %**	9,7 %

TABLE VI.8 – Nature des mots hors-vocabulaire corrigés selon le domaine d'application

des mots hors-vocabulaire) pour les données générales. Lors d'une tâche de traduction, des derniers doivent simplement être recopiés dans le résultat, tandis que les noms communs doivent être correctement traduits par leurs équivalents dans la langue cible (notre système de traduction de référence traite les mots hors-vocabulaire rencontrés en les recopiant systématiquement dans le résultat de traduction).

Il faut retenir de cette étude est que la post-édition statistique corrige 42 % des noms communs contenus dans les données spécialisées contre seulement 1 % pour ceux contenus dans les données générales (table VI.7).

L'analyse de la correction des mots hors-vocabulaire sur le corpus de spécialité nous a permis de constater que le SPES permet de corriger des mots ou des termes très spécifiques au domaine et qui apparaissent de façon fréquente dans les données (les fréquences sont données entre parenthèses), comme par exemple : ions (x 54), évaporites (x 39), électrolytes (x 26), subsurface (x 17), saumures (x 8) et tubirdité (x 5).

Nos résultats expérimentaux montrent donc que, lorsqu'il est appliqué à des données spécifiques à un domaine, notre système de post-édition statistique corrige un grand nombre de noms communs hors-vocabulaire. Cela explique, en partie, l'amélioration globale de la qualité de la traduction constatée sur le score TER.

Pour résumer : les systèmes de post-édition statistiques ne semblent pas permettre, au jour d'aujourd'hui et avec leur architecture actuelle, de corriger efficacement les résultats d'un PBMT du domaine général, mais ils permettraient d'effectuer des tâches d'adaptation au domaine grâce à leur faculté à restaurer du vocabulaire spécifique au domaine. Ce résultat fait émerger la question suivante : comment se situent la post-édition statistique par rapport aux autres techniques traditionnellement employées pour des tâches d'adaptation au domaine ?

Comparaison avec d'autres techniques d'adaptation au domaine

La post-édition statistique semblant être une technique efficace pour la tâche d'adaptation au domaine, nous proposons de comparer cette approche à d'autres méthodes traditionnellement utilisées pour l'adaptation au domaine. Pour cela, nous considèrerons deux méthodes « usuelles » d'adaptation au domaine, l'une basée sur le corpus et l'autre sur les modèles.

La première méthode d'adaptation au domaine étudiée est celle qui consiste simplement à concaténer le corpus spécifique au domaine, au corpus général d'apprentissage avant d'entraîner le système de traduction. Comme il est possible de le constater dans

les résultats présentés dans le tableau VI.9 ligne (2), cette méthode naïve permet d'obtenir un gain significatif en termes de scores BLEU et TER (+37,0 % et -25 %) en dépit du fait que les données de type spécifiques ne représentent que 0,5 % du corpus total d'apprentissage. Néanmoins, ce gain augmente si on augmente le poids attribué aux données spécifiques dans le corpus d'apprentissage, ici par duplication des données (résultats lignes (3), (4) et (5)). Le système atteint sa meilleure performance en termes de scores BLEU et TER (+48,2 % et -45,0 %) avec des données spécifiques représentant 35,5 % de la taille totale du corpus d'apprentissage (ligne (4)).

Systèmes	TER (*BLEU*)
PBMT Général	*46,7 (33,3)*
(1) SPES et domaine spécifique	39,2 (*40,1*)
—— Adaptation basée sur le corpus ——	
(2) 1×corpus du domaine spécifique (= 0,5 %)	35,2 (*45,5*)
(3) 10×corpus du domaine spécifique (= 5,2 %)	33,1 (*48,5*)
(4) 10^2×corpus du domaine spécifique (= 35,5 %)	32,3 (*49,2*)
(5) 10^3×corpus du domaine spécifique (= 84,5 %)	32,6 (*48,9*)
—— Adaptation basée sur les modèles ——	
(6) $TT_s + TT_g + ML_s + ML_g$	33,0 (*47,9*)
(7) $TT_s + TT_g + ML_i$	32,2 (*49,2*)

TABLE VI.9 – Performances — scores TER (*BLEU*) — sur un domaine spécifique selon la méthode d'adaptation au domaine

Les méthodes d'adaptation au domaine basées sur les corpus nécessitent de réapprendre entièrement un système de traduction et augmentent considérablement la taille du corpus d'apprentissage (et donc la durée nécessaire à celui-ci). Par la suite, plutôt que de concaténer toutes les données d'apprentissage disponibles, nous avons testé deux méthodes basées sur les tables de traduction (TT) multiples et des modèles de language (ML).

Dans un premier temps, nous avons construit des tables de traductions et des modèles de langage pour chacun des ensembles de données (TT_s et ML_s spécifiques au domaine et, TT_g et ML_g de domaine général). Dans un deuxième temps, nous les avons combinés dans le modèle log-linéaire du système de traduction. Le résultat en termes de scores BLEU et TER de cette méthode d'adaptation, nommée « $TT_s + TT_g + ML_s + ML_g$ », apparait ligne(6) du tableau VI.9.

Nous avons ensuite testé une deuxième technique consistant à interpoler les deux modèles de langage, général et spécifique (respectivement ML_g et ML_s), en un seul modèle de langage (appelé par la suite ML_i). Les poids d'interpolation des modèles de langage sont estimés en utilisant l'algorithme EM [47] (pour *Expectation-Maximisation* en anglais) puis les deux modèles sont fusionnés (en utilisant l'outil SRILM [Stolcke 2002]) dans un modèle unique. Nous observons une légère amélioration en termes de

47. http ://sourceforge.net/apps/mediawiki/irstlm/index.php ?Title=LM_interpolation

Hypothèse de traduction	...avec adaptation du SPES (système (1) table VI.9)	...avec adaptation TT_s+TT_g+ ML_i (système (7) table VI.9)
• There is some maximum quantity of water vapor for each of the value of the air temperatures.	• There is **some maximum** amount of water vapor for **each of the** value of the air temperature.	• There is **a certain** amount of water vapor **maximum possible** for **every** value of the air temperature.
• This is in connection with the effects of noise.	• This is in **connection with the effects of acoustic**.	• This is in **relation to the acoustic effects**.
• A reduction in consumption of animal products will very probably a positive effect on consumption of water to agriculture	• A **shift** in consumption of **animal products** will **most likely** positive effect on water consumption **to** agriculture	• A **reduction** in **the** consumption of **products of animal origin** will **very probably** a positive effect on water consumption **of** agriculture

TABLE VI.10 – Exemples de traduction selon la méthode d'adaptation au domaine

gains absolus BLEU et TER avec cette méthode (dénommée « TT_s + TT_g + ML_i », ligne(7) dans le tableau VI.9).

Les résultats montrent que les techniques « usuelles » d'adaptation au domaine présentées dans les expérimentations (scores TER de 32,2 à 35,2) permettent de surpasser nettement le résultat obtenu en mettant en oeuvre un système de post-édition statistique (TER de 39,2). La table VI.10 montre quelques exemples de traduction adaptées au domaine en utilisant un système de post-édition appris sur des données spécialisées ou en utilisant la meilleure méthode d'adaptation au domaine (« TT_s + TT_g + ML_i »).

4 Conclusion

A travers ces expériences, nous avons tenté de fournir une meilleure compréhension de l'utilité des systèmes de post-édition statistique dans le but d'améliorer les résultats d'un système de traduction probabiliste. Pour cela, nous avons tenté de répondre aux questions suivantes : est-ce que les résultats des SPES appris sur des données « simulées » sont comparables à ceux appris sur des données « réelles » ? Est-ce qu'un SPES peut améliorer les résultats d'un système de traduction probabiliste généraliste ? Est-ce qu'un SPES peut être utilisé pour adapter un système de traduction généraliste vers un domaine spécifique ? Dans ce dernier cas, est-ce que les SPES sont plus efficaces que les méthodes « usuelles » d'adaptation au domaine ?

En premier lieu, nous avons montré qu'un système de post-édition statistique appris sur des données de domaine général de taille moyenne ($\approx$ 9 000 phrases) n'apporte aucun gain en terme de score BLEU et TER quand il est appliqué sur les résultats d'un système de traduction probabiliste général. Avec de tels paramètres expérimen-

taux, l'utilisation d'hypothèses de traduction corrigées manuellement (configuration « réelle ») à la place de traductions professionnelles indépendantes du système (configuration « simulée ») donne des résultats légèrement plus satisfaisants.

Nous avons également observé que l'ajout de données pour l'apprentissage du SPES n'est pas suffisant pour améliorer significativement les performances du système de traduction de référence. Quelle que soit la taille des données disponibles pour l'apprentissage sur SPES, il semble difficile d'améliorer et de corriger les résultats d'un système de traduction probabiliste généraliste à l'aide d'une post-édition statistique.

Cependant, en comparant notre SPES appris sur un domaine général à un SPES appris sur un domaine spécifique, nos expériences montrent que ce dernier nous permet d'atteindre des performances significativement meilleures. Un SPES est donc plus efficace quand il est appris sur un corpus de données issues d'un domaine spécifique et, dans ce contexte, il peut être utilisé avec succès pour adapter un système de traduction probabiliste généraliste à un domaine spécifique. Nous avons remarqué que ce gain de qualité était dû, en grande partie, au fort taux de noms communs hors-vocabulaire du système général, reconnus et corrigés par le SPES spéficique au domaine.

Dans nos dernières expériences, nous avons choisi de comparer l'adaptation au domaine faite par le système de post-édition statistique à d'autres méthodes d'adaptation au domaine qui consistent à apprendre des modèles de traduction et des modèles de langage spécifiques au domaine puis à les interpoler avec les modèles généraux. Dans notre protocole, les méthodes d'adaptation utilisées partagent le même système de traduction de référence et les même corpus d'apprentissage et de test. La qualité des résultats des approches basées sur l'adaptation des modèles de langage et de traduction dépasse celle des résultats obtenus avec le SPES spécifique. Il est toutefois important de noter que ces méthodes d'adaptation au domaine nécessitent l'apprentissage d'un nouveau système de traduction probabiliste. On peut facilement imaginer des situations pratiques où il serait impossible de créer un nouveau système (c'est le cas d'un système de traduction utilisé comme une « boîte noire ») ou des situations où il serait plus avantageux de conserver un système de traduction généraliste sur lequel seraient appliqués plusieurs SPES adaptés à différents domaines.

Chapitre VII

Utilisation du corpus de post-éditions pour l'estimation de confiance en traduction automatique

Malgré les progrès récents, les technologies de traduction automatique n'ont pas encore atteint une maturité suffisante pour garantir une performance comparable à celle d'experts humains. La tendance de l'estimation de confiance est apparue avec le besoin d'évaluer et mesurer la fiabilité que l'on peut accorder à une hypothèse de traduction fournie par un système automatique.

Cette problématique, fortement liée à la tâche d'évaluation humaine, nous a semblé être un terrain d'expérimentation intéressant pour exploiter notre corpus de post-éditions présenté dans la partie IV.

1 Principe des mesures de confiance

1.1 Qualité d'une traduction automatique et productivité en terme de post-édition

L'utilisation grandissante des résultats de systèmes de traduction automatique comme « pré-traduction » fait émerger un contexte où la qualité d'une traduction automatique est estimée en fonction des post-traitements nécessaires pour la transformer manuellement en une « bonne » traduction. Le rapport entre le résultat obtenu et les ressources mises en œuvre pour l'obtenir est appelé « productivité » et peut être mesuré par le coût de ce post-traitement ou « l'effort de post-édition ». Ainsi, moins une traduction automatique nécessitera d'opérations de post-édition, plus sa qualité sera estimée comme grande.

La corrélation entre les métriques d'évaluation de la qualité et la productivité en termes de post-édition est calculée soit à partir d'informations réelles provenant de l'action de post-édition (nombres d'opérations effectuées, temps passé, mesures oculométriques, etc.) soit sur la base d'estimation subjective de l'effort de post-édition faite par des annotateurs. Bien souvent, c'est cette dernière solution qui est retenue

car moins coûteuse et ne nécessitant pas de post-édition de la traduction. Les échelles de mesure se basent alors sur le taux de post-édition estimé par des annotateurs qui évaluent l'effort nécessaire pour transformer une traduction automatique donnée en une traduction de qualité « publiable ». Les échelles comportent généralement 3, 4 ou 5 classes ordonnées représentant des niveaux de qualité. La première classe représente généralement les traductions automatiques les plus « mauvaises », pour lesquelles il est plus productif de re-traduire directement à partir de la phrase source, et la dernière classe représente les traductions automatiques de bonne qualité, publiables telles quelles sans post-édition nécessaire. Les consignes de classification contienent traditionnellement des descriptions de classes de la forme [Specia et al. 2009b] :

1. une re-traduction complète est requise ;
2. la post-édition est plus rapide qu'une re-traduction ;
3. peu d'opérations de post-édition sont nécessaires ;
4. convient telle quelle pour le but recherché.

Il est toutefois important de noter que la notion d'effort de post-édition dépend des mesures utilisées pour la calculer. L'effort de post-édition tel que perçu par un annotateur humain, est un score subjectif qui peut être influencé par des biais de perceptions. Dans [Koponen 2012] par exemple, les auteurs comparent l'effort cognitif, ou l'effort de post-édition évalué par les annotateurs humains, avec l'effort « technique » ou le nombre réel d'opérations d'édition effectuées. Les résultats expérimentaux soulignent le fait que ces deux mesures ne sont pas toujours corrélées et que, par exemple, les phrases longues (en termes de nombre de mots) sont évaluées comme cognitivement coûteuses et ce, même si le taux de post-édition requis est relativement bas.

Par la suite, nous mesurerons la qualité d'une traduction fournie par un système automatique en fonction du taux de post-édition nécessaire pour la transformer en une traduction répondant aux critères donnés dans la partie IV.4. Ce taux de post-édition sera estimé par la métrique TER calculée entre les hypothèses de traduction automatiques et leurs post-éditions.

1.2 Définition de la mesure de confiance

L'estimation de confiance pour la traduction automatique permet d'estimer la qualité d'un résultat de traduction automatique en temps réel, sans avoir recours à une traduction de référence.

Les modèles actuels utilisent des techniques d'apprentissage automatique pour prédire un score de qualité à partir d'un ensemble d'indicateurs extraits de l'hypothèse de traduction à évaluer, de la phrase source dont elle est issue, et éventuellement d'un ensemble de ressources additionnelles. Etant donnés un segment source S, sa traduction automatique H et un ensemble de ressources additionnelles R, l'objectif des systèmes de mesures de confiance est d'utiliser des indicateurs de caractéristiques extraits du triplet *(S, H, R)*, pour apprendre un modèle capable de prédire un score représentant la confiance que l'on peut accorder à l'hypothèse de traduction H.

Cet axe de recherche est relativement récent puisque les premiers travaux sur les mesures de confiance pour la traduction automatique ont été présentés en 2003 dans

[Blatz et al. 2003]. L'intérêt s'est renforcé récemment avec le besoin croissant de développer des méthodes de détection d'erreurs et d'analyse des sorties de systèmes de traduction automatique comme par exemple dans [Mohit et Hwa 2007].

Même si les mesures de confiance portent généralement sur les phrases, quelques études se sont intéressées à l'estimation au niveau sous-phrasique (segments, mots).

Les axes de recherche actuels portent essentiellement sur les techniques d'apprentissage automatique utilisées et les indicateurs utiles dans l'évaluation de la qualité de la traduction.

1.3 Evaluation de systèmes de traduction et mesures de confiance

Les métriques automatiques usuelles commme BLEU, NIST ou TER produisent des scores de qualité en calculant la similarité entre une hypothèse issue d'un système de traduction automatique et une seconde traduction donnée comme référence. Ce type d'évaluation a pour but d'évaluer la capacité « globale » d'un système à traduire des textes et est, par exemple, utilisée pour comparer différents systèmes dans les campagnes d'évaluation. De part leur nature, ces métriques basées sur un calcul de similarité sont très controversées et nécessitent des traductions de référence coûteuses à produire.

En prenant en compte ces problèmes, plusieurs études se sont penchées sur des métriques permettant d'évaluer la qualité de la traduction à l'échelle de la phrase, sans avoir besoin de traduction de référence. L'idée est ici de produire une métrique « locale » ayant pour objectif d'évaluer la capacité du système à traduire une phrase particulière.

1.4 Usages des mesures de confiance

La recherche en estimation de confiance est motivée par le besoin de disposer d'une évaluation d'un résultat de traduction automatique, en temps réel, et sans nécessiter de traduction de référence.

L'estimation de confiance peut prouver son utilité dans tous les cas nécessitant une estimation de la qualité d'une sortie de système de traduction automatique et peut fournir une information précieuse, entre autres, pour :

- décider si une traduction donnée est de qualité suffisante pour être publiée telle quelle [Soricut et Echihabi 2010] ;
- informer des utilisateurs monolingues de la langue source sur la confiance à accorder à une traduction donnée ;
- filtrer les traductions qui ne sont pas assez bonnes pour que leurs post-éditions permettent un gain de temps au traducteur professionnel [Specia 2011, He et al. 2010] ;
- sélectionner la meilleure parmi plusieurs hypothèses de traduction provenant ou non de systèmes différents [Soricut et Narsale 2012].

2 Modèles de prédiction pour l'estimation de confiance

Les travaux actuels sur l'apprentissage automatique de modèles d'estimation de confiance s'intéressent principalement à deux caractéristiques essentielles : la méthode d'apprentissage utilisée et le type d'indicateurs qu'ils considèrent.

2.1 Méthodes d'apprentissage pour les modèles de prédiction

Les techniques d'apprentissage pour les modèles de prédiction reposent généralement sur des algorithmes utilisant des méthodes supervisées. Etant donné un ensemble de prédicteurs X donnés en entrée, le but est de prédire les valeurs des variables de sortie Y, à partir d'un ensemble d'apprentissage annoté (X, Y).

Le type d'approche utilisé dépend en partie de la nature des variables de sortie. La tâche d'apprentissage de mesures de confiance est formulée soit comme une tâche de classification (on cherche à prévoir les valeurs d'une variable discrète ou catégorielle), soit comme une tâche de régression (on cherche à prévoir les valeurs d'une variable continue), soit, finalement, comme une tâche de classement (on cherche à ordonner un ensemble d'hypothèses selon l'estimation de leurs qualités).

Par la suite, on s'intéressera aux approches par classification et régression.

Approche par régression

Les problèmes de régression consistent à prédire, à partir d'une base d'apprentissage, des valeurs numériques continues pour un ensemble de données.

Le but est d'apprendre une correspondance entre un vecteur d'entrée (constitué de valeurs d'indicateurs) de dimension n : $X = \mathbb{R}^n$, et un espace de valeurs (qui peuvent être bornées ou non) : $Y = \mathbb{R}$.

Dans le cas de la traduction automatique, l'approche par régression consiste à considérer la tâche d'estimation de confiance comme la prédiction d'un score continu représentant la qualité d'une phrase traduite [Albrecht et Hwa 2007, Specia et al. 2009a, Specia et al. 2009b].

Approche par classification

Les problèmes de classification sont un cas particulier de régression où les valeurs des variables à prédire sont discrètes. La classification consiste à prédire les classes d'un ensemble de données à partir d'une base d'apprentissage pré-classifiée. Le but est d'apprendre une correspondance entre un vecteur d'entrée (constitué de valeurs d'indicateurs) de dimension n : $X = \mathbb{R}^n$, et l'espace des classes de dimension m : $Y = \{C_1, C_2, ..., C_m\}$. Cette correspondance est établie par un classifieur.

Parmi les travaux reposant sur cette approche, on peut citer [Quirk 2004], [Kulesza et Shieber 2004] ou encore [Blatz et al. 2003] qui considèrent la tâche d'estimation de confiance comme une classification binaire des traductions automatiques en « bonnes » ou « mauvaises ».

Techniques d'apprentissage

Divers techniques d'apprentissage supervisé ont montré leur efficacité pour la modélisation d'estimateurs de confiance en traduction automatique : les arbres de décision [Soricut et al. 2012], les modèles de régression logistique [Raybaud et al. 2011], les réseaux de neurones [Gandrabur et Foster 2003], les modèles des moindres carrés partiels ou PLS (pour *Partial Least Square* en anglais) [Suzuki 2011] ou encore les Séparateurs à Vaste Marge ou SVM (pour *Support Vector Machine* en anglais)[Specia et al. 2009a].

Les SVMs ont été développés dans les années 1990 et représentent un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Il s'agit de la technique que nous avons sélectionnée pour modéliser notre système de prédiction.

2.2 Indicateurs de qualité pour les modèles de prédiction

L'apprentissage de mesures de confiance se fait sur la base de caractéristiques décrivant un ou plusieurs éléments du triplet : {phrase source, hypothèse de traduction, système de traduction}. On appelle indicateurs de qualité pour les modèles de prédiction, les variables caractéristiques qui vont permettre d'estimer la qualité de la traduction.

Ces variables peuvent être monolingues, bilingues, elles peuvent être dépendantes ou indépendantes du système de traduction et peuvent provenir en partie de ressources linguistiques externes, additionnelles.

Types de caractéristiques

Ces caractéristiques, qui ont pour objectif de fournir des informations sur la difficulté de traduction d'une phrase, peuvent prendre en considération divers éléments comme :

- les longueurs des phrases (la longueur de la phrase source, de la phrase cible, le rapport des longueurs, etc.) ;
- les probabilités de la phrase source et cible selon un modèle de langage ;
- les statistiques de fréquence des mots et n-grammes des phrases sources et cible (ratio type-token d'un mot de la phrase, la fréquence d'un n-gramme calculé sur un corpus monolingue, etc.) ;
- les signes de ponctuation et leurs nombres dans les phrases sources et cibles (appariement, superposition, etc.) ;
- les scores fournis par le système de traduction (les scores d'alignement, le score global de décodage, etc.) ;
- des ressources linguistiques additionnelles (WordNet, étiquettes morpho-syntaxiques, etc.).

Il est important de noter que si le type des caractéristiques peut paraître intuitif et leur utilisation évidente, trouver la forme exacte pour les introduire dans le système représente un travail conséquent.

Sélection d'indicateurs de qualité

Dans un premier temps, les études en estimation de confiance pour la traduction automatique se sont concentrées sur la découverte de nouvelles sources d'information pour les indicateurs de qualité. Les systèmes créés étaient alors appris sur plusieurs dizaines de variables caractéristiques comme par exemple dans [Blatz et al. 2003] où le système utilise 91 fonctions de caractéristiques.

Peu à peu, des travaux se sont intéressés à l'influence et à la sélection de ces indicateurs de qualité. Dans leur étude, [Specia et al. 2009a] constatent que la sélection d'indicateurs discriminants permet d'augmenter les performances du système : le meilleur résultat est obtenu avec une sélection de 32 indicateurs de qualité pertinents sur un total de 84.

Une manière simple d'étudier l'impact d'une variable caractéristique sur le score produit par le système d'estimation de confiance est, par exemple, de mesurer la corrélation entre ce score avec celui prédit par la fonction testée.

Indicateurs monolingues *vs* bilingues

Les variables caractéristiques utilisées pour l'estimation de confiance peuvent reposer uniquement sur l'information monolingue contenue dans la phrase source ou peuvent prendre en considération l'hypothèse fournie par le système de traduction (c'est-à-dire le couple {phrase source, hypothèse de traduction}).

Dans le premier cas, on parlera d'indicateurs monolingues. L'analyse de la phrase source peut être enrichie par l'utilisation de corpus représentatifs de la langue étudiée ou l'application d'outils d'analyse morphologique ou syntaxique. L'intérêt est ici de pouvoir prédire la difficulté de traduction de la phrase source en amont de tout processus de traduction.

Les exemples (1), (9), (10) et (11) du tableau VII.1 (voir page 123) sont, par exemple, des indicateurs monolingues alors que les exemples (3) (5) et (6) sont des indicateurs bilingues.

Indicateurs "black-box" et "glass-box"

Parmi les variables caractéristiques présentées dans la littérature, on peut distinguer celles extraites du processus de traduction (indicateurs "*glass-box*") et celles qui sont extraites des phrases sources et cibles (éventuellement d'autres corpus) mais qui sont indépendantes du processus de traduction (indicateurs "*black-box*").

L'intérêt de ces dernières est, d'une part, qu'elles peuvent être utilisées lorsque l'on n'a pas accès au système de traduction (ce qui peut être le cas pour des systèmes de traduction commerciaux), et d'autre part, qu'elles peuvent être appliquées à différents outils de traduction.

Même si les indicateurs de qualité indépendants du système de traduction contribuent de façon certaine à la performance des systèmes actuels (ils représentent, par exemple, plus de la moitié des indicateurs sélectionnés dans [Specia et al. 2009a]), à ce jour, aucune étude ne montre la possibilité d'envisager l'apprentissage d'un sys-

tème d'estimation de confiance efficace uniquement sur la base de ce type d'indicateurs ("*black-box*").

Les fonctions (14), (15) et (16) du tableau VII.1 (voir page 123) sont des exemples d'indicateurs dépendants du système de traduction.

3 Mesures de confiance apprises à partir de post-éditions

La partie suivante décrit l'élaboration et l'évaluation de modèles d'estimation de confiance appris sur notre corpus de post-éditions présenté dans la partie IV.

3.1 Cadre expérimental

Campagne d'évaluation WMT

Dans le souci de situer nos travaux par rapport à l'état de l'art des études existantes, nous avons choisi de sélectionner nos paramètres expérimentaux de façon à pouvoir comparer nos résultats avec ceux obtenus par les participants à la tâche d'estimation de qualité de la campagne d'évaluation internationale WMT ayant eu lieu pour la première fois, en juin 2012, à Montréal au Canada. Les organisateurs de la campagne d'évaluation proposent, pour la tâche, un cadre expérimental de référence inspiré des travaux présentés dans [Specia 2011]. Celui-ci contenant :

- un ensemble de 17 indicateurs de qualité ;
- un corpus constitué de phrases source anglaises et de leurs traductions en espagnol faites par un système de traduction statistique basé sur Moses. Ce corpus est décomposé en un ensemble de 1 832 phrases destinées à l'apprentissage du système d'estimation de confiance et 442 phrases destinées à son évaluation ;
- un score de qualité associé à chaque hypothèse de traduction du corpus et calculé sur la base de l'effort de post-édition estimé par trois annotateurs. Le score est continu dans l'intervalle [1; 5] et représente la moyenne des 3 scores proposés par les annotateurs en respectant les consignes fournies dans le tableau de la figure VII.1 ;
- deux métriques d'évaluation du modèle de prédiction : MAE et RMSE (présentées par la suite dans la partie 3.1.0) ;

Le détail des paramètres expérimentaux et les résultats des participants à la campagne d'évaluation sont présentés dans [Callison-Burch et al. 2012].

Il est important de noter que même si le système d'évaluation de référence proposé est considéré comme une « baseline », il représente en réalité un système performant qui a prouvé son efficacité pour plusieurs paires de langues, systèmes de traduction et corpus d'application : lors de l'évaluation de 2012, seuls 5 participants sur 20 sont parvenus à améliorer significativement les résultats obtenus par celui-ci.

Les expériences proposées, par la suite, dans ce chapitre s'inspirent en grande partie du protocole expérimental de la campagne d'évaluation WMT 2012. La principale différence est la substitution du corpus fourni lors de la campagne (corpus anglais/espagnol

de 2 274 phrases) par notre corpus décrit dans la partie IV (corpus français/anglais de 10 881 phrases) et l'utilisation de taux de post-édition calculés de manière empirique (sur des post-éditions humaines) plutôt qu'estimés de manière subjective par des annotateurs.

Classe	Taux de post-édition (PE)
1	≥ 70 % de PE
2	50 % $\geq$ PE $\leq$ 70 %
3	25 % $\geq$ PE $\leq$ 50 %
4	10 % $\geq$ PE $\leq$ 25 %
5	≤ 10 % de PE

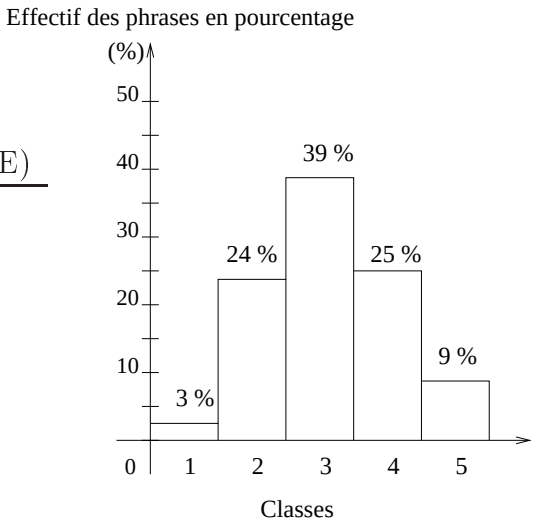


FIGURE VII.1 – Consignes de l'annotation de données WMT'12 : classement des traductions automatiques selon le taux de post-édition estimé par les annotateurs et répartition des 1832 phrases du corpus

Indicateurs de qualité sélectionnés

Le corpus considéré est celui décrit dans la partie IV : il contient 10 881 phrases sources françaises traduites en anglais par notre système de traduction de référence.

Les indicateurs retenus pour nos expérimentations sont les 17 fonctions de caractéristiques présentées dans le système de mesure de qualité de référence de la campagne d'évaluation WMT 2012, auxquels nous avons ajouté un indicateur représentant le nombre de noms communs de la phrase source inconnus du modèle de traduction.

Les 18 indicateurs de qualité utilisés pour notre modèle d'estimation de confiance sont listés ci-après :

- le nombre de mots dans la phrase source ;
- le nombre de mots dans l'hypothèse de traduction ;
- le nombre moyen de caractères dans les mots de la phrase source ;
- la probabilité de la phrase source selon un modèle de langage source ;
- la probabilité de l'hypothèse de traduction selon un modèle de langage cible ;
- le nombre moyen d'occurrences des mots dans l'hypothèse de traduction (ratio type/token) ;
- le nombre moyen de traductions par mot de la phrase source donné par le modèle de traduction (on considère seulement les traductions pour lesquelles la probabilité de traduction lexicale est strictement supérieure à 0,2) ;
- le nombre moyen de traductions par mot de la phrase source donné par le modèle de traduction (on considère seulement les traductions pour lesquelles la proba-

bilité de traduction lexicale est strictement supérieure à 0,01) pondéré par la fréquence inverse de chaque mot source dans le corpus d'apprentissage ;
- la proportion d'uni-grammes se situant dans le premier quartile (Q1) des fréquences d'uni-grammes du corpus source d'apprentissage du SMT [48] ;
- la proportion d'uni-grammes se situant dans le dernier quartile (Q4) des fréquences d'uni-grammes du corpus source d'apprentissage du SMT ;
- la proportion de bi-grammes se situant dans le premier quartile (Q1) des fréquences de bi-grammes du corpus source d'apprentissage du SMT ;
- la proportion de bi-grammes se situant dans le dernier quartile (Q4) des fréquences de bi-grammes du corpus source d'apprentissage du SMT ;
- la proportion de tri-grammes se situant dans le premier quartile (Q1) des fréquences de tri-grammes du corpus source d'apprentissage du SMT ;
- la proportion de tri-grammes se situant dans le dernier quartile (Q4) des fréquences de tri-grammes du corpus source d'apprentissage du SMT ;
- la proportion des uni-grammes de la phrase source présents dans le corpus source d'apprentissage du SMT ;
- le nombre de signes de ponctuation dans la phrase source ;
- le nombre de signes de ponctuation dans l'hypothèse de traduction ;
- le nombre de noms communs de la phrase source inconnus par le modèle de traduction.

A chaque phrase du corpus est donc associé un ensemble de 18 valeurs correspondants aux indicateurs de qualité présentées ci-dessus. Ces valeurs sont, par la suite, normalisées dans l'intervalle $[-1; 1]$.

Chaque phrase du corpus est donc représentée par un vecteur $x_i \in [-1; 1]^m$ avec m le nombre de prédicteurs (ici $m = 18$) et $i \in 1, ..., n$, n étant la taille du corpus en nombre de phrases (ici $n = 10881$).

Lors des expérimentations, le corpus de 10 881 phrases est divisé en un ensemble d'apprentissage et un ensemble de test disjoints.

Apprentissage par Séparateurs à Vastes Marges

Le corpus d'apprentissage annoté est utilisé pour entrainer des Séparateurs à Vaste Marge (ou SVM) à l'aide du logiciel LIBSVM [Chang et Lin 2011]. Les séparateurs appris utilisent une fonction noyau à base radiale (ou RBF pour *radial basis function* en anglais). Il est à noter que d'autres fonctions ont été testées pour des résultats similaires à ceux obtenus avec le noyau RBF. Les valeurs des deux paramètres de la fonctions noyau (γ et C) sont ajustés par une technique de validation croisée sur le corpus d'apprentissage (partition de 5 sous-échantillons).

La technique SVM est utilisée pour apprendre, d'une part, un modèle de régression destiné à prédire la valeur numérique du score TER associé à un couple *(phrase source,*

48. Calcul de la fréquence des n-grammes contenus dans le corpus d'apprentissage source puis utilisation des quartiles (Q1,Q2,Q3,Q4) de cet ensemble pour calculer les appartenances des n-grammes de la phrase source à ces classes. On rappelle que la premier quartile contient les mots du corpus les plus rares (les moins fréquents) alors que le dernier quartile contient les mots les plus fréquents du corpus.

hypothèse de traduction) donné, et d'autre part, un classifieur destiné à associer une classe de qualité à ce même couple.

Evaluation

L'intérêt prédictif des modèles appris est évalué sur un corpus de test de 881 phrases. Les valeurs prédites par les modèles et les valeurs de référence associées aux phrases sont utilisées pour calculer deux indicateurs statistiques de précision : la moyenne arithmétique des valeurs absolues des écarts entre les prévisions et les références (ou MAE pour *Mean Absolute Error* en anglais) et l'erreur quadratique moyenne ou l'erreur-type (ou RMSE pour *Root Mean Squared Error* en anglais), la racine carrée de la moyenne des différences au carré entre les prévisions et les références dont les formules sont les suivantes :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

avec $N = 881$, $\hat{y}_i$ la valeur prédite par le modèle et y_i la valeur de référence du corpus.

Les scores MAE et RMSE sont des mesures de même unité que les données : un score MAE de 0,80 signifiera qu'en moyenne, la différence entre la valeur de référence et la valeur de la prédiction du modèle est de 0,80 points.

Les scores MAE et RMSE sont continus dans l'intervalle $[0; +\inf[$. Plus ils sont faibles, meilleur est le modèle de prédiction évalué.

L'interprétation du score RMSE est similaire à celle du score MAE mais ce dernier est plus sensible aux erreurs de prévisions les plus importantes : lors du calcul de la métrique RMSE, les écarts entre les valeurs prédites et les valeurs de référence sont élevés au carré avant d'être moyennés, ce qui a pour résultat de donner un poids important aux grandes erreurs d'estimation.

3.2 Modèles d'estimation de scores TER

On s'intéresse ici à l'estimation du score TER tel que calculé entre l'hypothèse de traduction et sa post-édition humaine.

Chaque phrase du corpus est représentée par :

- un vecteur $x_i \in [-1; 1]^n$ avec n le nombre d'indicateurs (ici $n = 18$) et $i \in 1, ..., l$, l étant la taille du corpus en nombre de phrases ;
- une valeur $y_i \in [0; +\inf[$ représentant le score TER de référence calculé entre l'hypothèse de traduction et sa post-édition.

Ce problème de régression est traité par une fonction de prédiction apprise sur le corpus d'apprentissage, qui a chaque couple $(x_i; y_i)$ associe une valeur $\hat{y}_i \in [0; +\inf[$ représentant le score TER prédit par le modèle.

Deux systèmes de prédiction sont appris : le premier sur un corpus d'apprentissage de 1 000 phrases (système (1)) et le second sur un corpus d'apprentissage de 10 000

phrases (système (2)). L'évaluation des deux systèmes se fait sur un même corpus de test de 881 phrases.

Système	Taille du corpus d'apprentissage	MAE	RMSE
(1)	1 000 phrases	0,14	0,19
(2)	10 000 phrases	0,13	0,17

TABLE VII.2 – Résultats des systèmes d'estimation du score TER sur les hypothèses de traductions automatiques

Les résultats donnés dans le tableau VII.2 montrent que le système (2) appris sur un corpus de 10 000 phrases obtient un score MAE de 0,13 ce qui signifie que le modèle d'estimation est capable de prédire un score TER avec, en moyenne, un écart de 0,13 points comparé au score TER de référence calculé entre l'hypothèse de traduction et sa post-édition. Le modèle appris estime donc avec une précision relativement importante (en moyenne 0,13 points d'erreur) le taux de post-édition d'une hypothèse de traduction exprimé en score TER.

Les résultats permettent également de constater que la taille du corpus d'apprentissage n'influe que très peu sur la qualité du prédicteur en termes de mesures MAE et RMSE : l'ajout de données d'apprentissage (de 1 000 phrases à 10 000 phrases), possible grâce à notre corpus, améliore le score MAE de 0,01 point et le score RMSE de 0,02 points.

3.3 Modèles de classification des hypothèses de traduction

L'objectif est ici de pouvoir associer automatiquement une hypothèse de traduction à une classe représentant sa qualité. Pour cela, nous avons choisi, sur le modèle de la campagne d'évaluation WMT, de définir cinq classes de qualité sur la base du taux de post-édition nécessaire pour parvenir à une « bonne » traduction de la phrase source : la classe 5 représentant les traductions jugées « bonnes », pour lesquelles les opérations de post-édition à effectuer sont peu nombreuses, voire nulles, et la classe 1 représentant les traductions jugées « mauvaises » pour lesquelles les opérations de post-édition sont estimées nombreuses.

Dans le souci de pouvoir comparer nos résultats à ceux présentés lors de la campagne d'évaluation WMT 2012, nous expérimentons deux répartition des données. La première répartition, présentée dans la figure VII.2, présente des intervales de classes cohérents avec les consignes données aux annotateurs lors de la classification des données de WMT : le score TER représentant, par sa définition, le taux de post-édition, les intervalles de classes des données WMT (voir figure VII.1) sont conservés tels quels.

Néanmoins, on remarque que les distributions entre la répartition faite à partir du taux de post-édition estimé sur les données WMT (voir figure VII.1) et celle faite à partir du taux de post-édition calculé empiriquement sur notre corpus (voir figure VII.2) ne sont pas comparables. Dans notre cas, la classification faite selon les consignes « théoriques » usuelles mène à une distribution hétérogène susceptible de rendre difficile la prédiction des classes à faible effectif.

Cette différence entre les distributions peut être expliquée par la nature des données : dans notre cas l'effort de post-édition est calculé automatiquement (*via* le score TER) entre l'hypothèse de traduction et sa post-édition alors que dans les données WMT le taux de post-édition est estimé de façon subjective par un annotateur humain.

Face à ce constat, nous proposons une deuxième répartition, décrite dans la figure VII.3 et présentant un diagramme de distribution s'approchant d'une courbe de Gauss (ou « en cloche »), comparable à celui des données de WMT.

Classe	Score TER
1	TER $\geq 0{,}70$
2	$0{,}50 \leq$ TER $< 0{,}70$
3	$0{,}25 \leq$ TER $< 0{,}50$
4	$0{,}10 \leq$ TER $< 0{,}25$
5	TER $\leq 0{,}10$

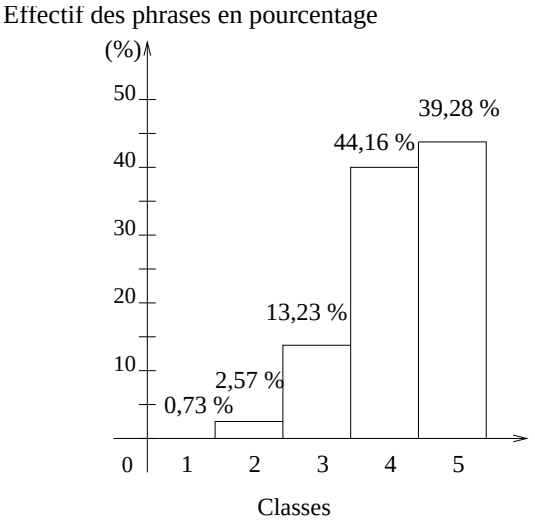


FIGURE VII.2 – Proposition de classement n° 1 selon le score TER et répartition des 10 881 phrases du corpus dans les classes

Classe	Score TER
1	TER $> 0{,}70$
2	$0{,}35 <$ TER $\leq 0{,}70$
3	$0{,}15 <$ TER $\leq 0{,}35$
4	$0 <$ TER $\leq 0{,}15$
5	TER $= 0$

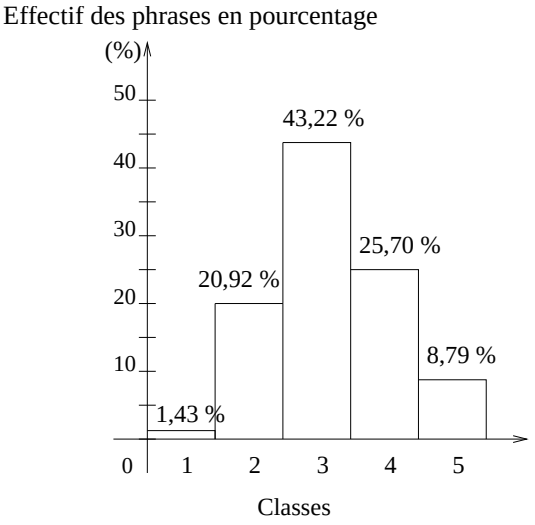


FIGURE VII.3 – Proposition de classement n° 2 selon le score TER et répartition des 10 881 phrases du corpus dans les classes

Après la distribution en classes, chaque phrase du corpus est représentée par :

- un vecteur $x_i \in [-1;1]^n$ avec n le nombre de prédicteurs (ici $n = 18$) et $i \in 1, ..., l$, l étant la taille du corpus en nombre de phrases ;
- une classe $c_i \in \{1, 2, 3, 4, 5\}$ représentant la qualité de l'hypothèse de traduction en termes de volume de post-édition à effectuer.

Ce problème de classification est traité par une fonction de prédiction, apprise sur le corpus d'apprentissage, qui a chaque couple $(x_i; c_i)$ associe une valeur $\hat{c_i} \in \{1, 2, 3, 4, 5\}$ représentant la classe prédite par le modèle.

Un système de prédiction est appris sur chacune des deux répartitions présentées précédemment. L'apprentissage des deux systèmes se fait sur un corpus de 10 000 phrases et l'évaluation sur un corpus de test de 881 phrases.

Système	Classes	Taux de précision	MAE	RMSE
(3)	Figure VII.2	48,5 % (428/881)	0,66	1,02
(4)	Figure VII.3	48,0 % (422/881)	0,63	0,93

TABLE VII.3 – Résultats des systèmes de prédiction de classe de qualité pour les hypothèses de traductions automatiques

Les résultats donnés dans le tableau VII.3 montrent que les deux systèmes obtiennent un taux de précision de plus de 48 % ce qui signifie que, près d'une fois sur deux, la classe prédite par le système d'estimation de confiance est identique à la classe donnée pour référence dans le corpus.

Le système (4), qui obtient de meilleures performances que le système (3), présente un score MAE de 0,63 et un score RMSE de 0,93. Si l'on interprète ces scores comme l'erreur moyenne du système, on peut conclure que le modèle de prédiction de qualité attribue des classes qui sont, en moyenne, très proches (dans le sens ordinal) des classes de référence.

Si l'on compare nos résultats à ceux obtenus par les participants à la tâche analogue de la campagne d'évaluation WMT 2012 [Callison-Burch et al. 2012], on constate que ceux-ci se situent à l'état de l'art voire parmi ceux des systèmes jugés comme meilleurs lors de l'évaluation. Cette comparaison doit cependant prendre en considération le fait que le critère de qualité diffère d'une expérimentation à l'autre : les scores de la campagne d'évaluation WMT sont calculés à partir d'efforts de post-édition estimés par des annotateurs humains et sont continus entre 1 et 5 (ils représentent une moyenne sur 3 annotateurs différents) alors que le critère utilisé dans nos expérimentation est discret et estimé à partir du taux de post-édition calculé entre une hypothèse de traduction et sa post-édition humaine.

4 Conclusion

Les résultats de l'étude présentée dans ce chapitre montrent que le corpus de post-éditions que nous avons collecté et présenté dans la partie IV, constitue un ensemble de données intéressant pour l'apprentissage de mesures de confiance pour la traduction automatique statistique.

Nous avons extrait un ensemble de 18 indicateurs de qualité de notre corpus de 10 881 phrases et nous avons utilisé les post-éditions collectées pour apprendre quatre Séparateurs à Vaste Marge (SVM) pour l'estimation de confiance.

Les résultats obtenus montrent que les estimateurs créés sur ces données produisent des modèles de qualité à l'état de l'art pour prédire le taux de post-édition nécessaire à une hypothèse de traduction donnée et/ou pour lui affecter une classe de qualité sur une échelle de 5 valeurs.

Source de l'information	Exemples de descripteurs
Nombre de mots ou caractères dans les phrases	1. Taille de S 2. Taille de H 3. Ratio entre la taille de S et H 4. Longueur moyenne des mots de S en caractères
Symboles numériques et signes de ponctuation des phrases S et H	5. Recouvrement des symboles numériques et signes de ponctuation entre S et H 6. Proportion et ratio des symboles numériques et signes de ponctuation dans S et H
Modèles de langage	5. Probabilité d'un modèle de langage n-gramme appliqué sur S et H
Modèle de traduction	5. Nombre moyen de traductions alternatives par mot de la phrase S pondéré ou non par la fréquence du mot
Distribution des n-grammes de S	5. Statistiques sur les fréquences de n-grammes dans le corpus d'apprentissage (= couverture de la phrase S dans le corpus d'apprentissage)
Analyse syntaxique et morphologique	5. Proportion de noms/verbes/adjectifs/etc. dans S 6. Probabilité de la séquence des étiquettes morpho-syntaxiques de S estimée par un modèle de langage
Liste de traductions candidates donnée par le décodeur (*n-best list*)	5. Ratio entre le nombre de mots différents dans la liste et la taille moyenne de la traduction (= uniformité des traductions de la liste) 6. Probabilité de H donnée par un modèle de langage appris sur la liste des traductions de la liste
Scores donnés par le décodeur du système de traduction (Moses)	5. Scores des fonctions de caractéristiques du modèle 6. Score global de traduction donné par le décodeur 7. Proportion de noeuds avortés dans le graphe de recherche du décodeur

TABLE VII.1 – Exemples d'indicateurs de qualité pour l'estimation de confiance en traduction automatique (avec "H" une hypothèse de traduction et "S" la phrase source dont elle est issue)

Conclusion et perspectives

Cette thèse présente nos contributions à l'étude de l'apport de corrections de résultats de traductions (ou post-éditions) pour améliorer un système de traduction automatique probabiliste.

Les travaux que nous avons présentés dans ce manuscrit sont issus d'une problématique et de motivations que nous pensons très actuelles dans le domaine de la traduction automatique. Après les avoir brièvement rappelés, je propose une vue synthétique des principaux résultats à retenir pour enfin discuter, dans une dernière partie, des perspectives de recherche qui s'en dégage.

Problématique et motivations

Les technologies de traduction automatique existantes sont à présent vues comme une approche prometteuse pour aider à produire des traductions de qualité de façon efficace et à coût réduit. Cependant, l'état de l'art actuel des systèmes de traduction automatiques est encore bien loin de permettre une automatisation complète du processus et la coopération homme/machine reste indispensable pour produire des résultats de qualité. Même si cette intervention humaine peut se manifester à plusieurs niveaux du processus de traduction, la plus usuelle consiste à post-éditer les sorties fournies par le système de traduction automatique, c'est-à-dire effectuer une vérification (et, si nécessaire, une correction) des résultats de traduction.

Ce travail de vérification et de correction effectué par les utilisateurs sur les résultats de traduction automatique constitue une source de données précieuses pour l'analyse et l'adaptation des systèmes.

Cependant peu de travaux expérimentaux étudient la faisabilité et la performance de l'intégration de post-éditions comme « corrections » du système automatique ayant généré les traductions. En effet, à ce jour aucun travail de recherche public ne mentionne de façon explicite comment un système de traduction automatique pourrait s'adapter pour bénéficier de retours donnés par des utilisateurs.

Les travaux présentés dans ce manuscrit visent à étudier l'apport de traductions post-éditées pour améliorer le système de traduction automatique et proposer différentes techniques pour exploiter les suggestions de corrections d'utilisateurs afin de tenter d'ajuster, au mieux, les systèmes automatiques.

Principales contributions

Nos expériences préliminaires ont montré le potentiel de l'approche mais également fait ressortir le besoin de collecter un corpus ayant les caractéristiques adéquates pour l'étude visée et les objectifs que nous nous sommes fixés. Nous avons donc collecté un corpus d'environ 10 000 énoncés français traduits en anglais par notre système de traduction automatique probabiliste puis post-édités par des annotateur volontaires et non nécessairement traducteurs professionnels. Le corpus collecté nous a permis d'analyser la spécificité de la nature d'un tel corpus mais aussi d'étudier plusieurs aspects liés à l'intégration de ces post-éditions au système de traduction. Les résultats de nos études sont détaillés ci-après.

Collecte et évaluation d'un corpus de post-éditions faites par des annotateurs volontaires et non nécessairement bilingues

Nous avons collecté un corpus de 12 381 hypothèses de traduction (soit environ 300 000 mots) post-éditées par des annotateurs volontaires non experts. Le corpus post-édité est composé de 10 881 hypothèses issues de notre système statistique de référence et 1500 traductions professionnelles utilisées usuellement comme référence pour l'apprentissage et le test des systèmes. Toutes sont des traductions anglaises d'énoncés français provenant de divers sites Web journalistiques.

Afin de limiter les coûts relatifs à la collecte, nous avons choisi d'utiliser un plateforme de *crowdsourcing*. Les problèmes légaux, économiques et éthiques impliqués par la méthode de collecte nous ont amenés à définir et respecter des principes de « bonne conduite » pour l'utilisation de l'outil. Le contrôle de la fiabilité du contenu créé par les annotateurs non experts a été réalisé par nos soins, *via* une vérification rigoureuse et systématique des annotations collectées. Le profil exigé des participants était une compréhension aisée de la langue francaise et une pratique courante de la langue anglaise sans nécessairement être natif anglophone. La collecte s'est déroulée sur une durée de 4 mois et 12 jours et a coûté au total 2 040 dollars. Près de 560 personnes inscrites sur Amazon Mechanical Turk ont participé à la tâche de post-édition.

Nous avons procédé à une évaluation de la qualité des énoncés collectés sur un sous-ensemble de 311 phrases du corpus post-édité *via* Amazon Mechanical Turk. Les résultats montrent que 81,35 % sont des traductions irréprochables ou acceptables de la phrase source et seulement 2,57 % contiennent une ou plusieurs erreurs qui font que la traduction reste fausse ou incomplète. La qualité des post-éditions collectées via Amazon Mechanical Turk a également été comparée, sur un échantillon de 111 phrases, avec des post-éditions faites par des traducteurs professionnels. Les résultats montrent que 93,7 % des post-éditions non-professionnelles que nous avons collectées sont considérées comme étant, au moins, de qualité professionelle.

Cette étude nous permet donc d'inférer que, dans le contexte des instructions données et des moyens mis en œuvre pour cette collecte, le corpus de post-éditions que nous avons collecté est de qualité quasi « professionnelle ». Les données recueillies sont mises gratuitement à la disposition de la communauté scientifique et leur téléchargement est possible, en ligne, à l'URL suivante :

http://www-clips.imag.fr/geod/User/marion.potet/index.php?page=download

Analyse du corpus de post-éditions d'hypothèses de traductions automatiques

Le corpus collecté contient 10 881 post-éditions d'hypothèses de traduction issues de notre système de traduction statistique mais aussi 1 500 post-éditions de traductions dites « de référence », mises à disposition de la communauté par le biais des corpus bilingues alignés. Ces dernières données (1 500 post-éditions) représentent des traductions faites par des traducteurs professionnels, souvent à l'échelle d'un document, indépendamment du système de traduction automatique pour lesquelle elles sont utilisées et sans l'objectif d'être ré-utilisées pour effectuer des tâches d'apprentissage automatique.

L'analyse des post-éditions collectées montre que 9 % des hypothèses de traduction ont été jugées par les post-éditeurs comme ne nécessitant pas de correction lors de la post-édition (c'est-à-dire que 9 % des résultats de notre système de traduction de référence sont considérés comme de parfaites traductions de la phrase source) alors que les même statistiques faites sur les traductions professionnelles de référence montrent que seulement 28 % d'entre elles sont condidérées comme correctes. Autrement dit, 72 % des traductions données comme référence dans les corpus bilingues ont nécessité une correction lors de la post-édition.

Ceci peut être expliqué par le fait que les traductions professionelles de référence fournies avec les corpus bilingues alignés sont, dans le cas des corpus utilisés ici, produites dans un contexte de traduction à l'échelle du texte dans son intégralité et non de la phrase en tant qu'unité indépendante. Ces dernières ne sont donc pas appropriées pour être considérées au niveau des phrases et de surcroît dans un contexte d'apprentissage automatique phrase-à-phrase d'un système de traduction probabiliste.

Dans un deuxième temps, nous avons utilisé le corpus collecté pour mesurer la similarité entre les différents types de traductions à l'aide des métriques WER, TER et d ($d = 100 - scoreBLEU$). Nous avons observé que la distance entre les traductions de référence et les hypothèses de traduction est deux fois plus importante que celle qui relie ces dernières à leurs corrections. De la même façon, les hypothèses de traductions corrigées et les traduction de référence sont excessivement éloignées alors que toutes deux sont censées être des traductions « correctes » des mêmes phrases sources.

Le constat qui ressort de ces travaux est que les post-éditions, qui sont des traductions correctes, en théorie aussi proches que possible de celles faites par le système de réference, sont plus propices à être utilisées dans un but d'adaptation voire de correction du système que les traductions faites à l'échelle d'un document, indépendamment du système de référence.

Exploitation du corpus collecté pour améliorer le système de traduction de référence

Nous avons mené une série d'expériences visant à intégrer le corpus de post-éditions précédemment collecté dans le système de traduction, dans le but de l'améliorer. Après avoir partitionné le corpus collectée en trois sous-ensembles (pour l'apprentissage, le

développement et le test des systèmes), nous avons évalué deux protocoles expérimentaux : l'enrichissement du système de traduction de référence par ajout d'une table de traduction apprise sur les post-éditions et la post-édition automatique des résultats du système de référence.

La première approche, qui consiste à ré-intégrer les post-éditions en enrichissant le modèle de traduction du système de référence avec une table de traduction apprise sur les post-éditions collectées, ne montre pas d'amélioration des scores automatiques par rapport au système de référence. Les résultats de la deuxième approche qui consiste à post-éditer automatiquement les résultats de traduction du système de référence, montrent que celle-ci constitue une piste de travail intéressante pour ré-intégrer les post-éditions humaines dans le processus de traduction. Même si l'usage d'un post-éditeur automatique « naïf » dégrade significativement la qualité des sorties « brutes » du système de traduction de référence, nous avons expérimenté plusieurs pistes d'adaptation du système afin de mieux évaluer le potentiel et les limites de l'approche.

Nous abservons dans un premier temps que l'ajustement des poids du modèle de post-édition à l'aide de la méthode MERT permet bien d'augmenter significativement la qualité des résultats sur le corpus de test et ce, quelque soit le type de référence utilisé (les traductions professionnelles indépendantes ou les post-éditions). Nous constatons également une amélioration des métriques d'évaluation automatique avec la méthode de filtrage de la table de traduction du système de post-édition statistique (gain de 3,5 % de score BLEU et 6,2 % de score TER par rapport au système de post-édition de référence) ainsi que le modèle de post-édition à architecture hiérarchique (gain de 3,4 % de score BLEU et 5,8 % de score TER par rapport au système de post-édition de référence). D'un autre côté, l'ajout d'un modèle de langage appris sur les post-édition et l'ajout du contexte source de traduction ne permettent pas d'outrepasser les performances du système de post-édition de référence.

Au delà de l'application systématique de la post-édition statistique, nous expérimentons la post-édition sélective des hypothèses de traduction. Bien que le score « oracle » (sélection des phrases *a posteriori*) calculé pour évaluer le potentiel de la méthode de post-édition sélective soit significativement meilleur que celui obtenu avec la post-édition systématique, les résultats montrent que le gain à espérer en développant une méthode de sélection automatique des phrases à post-éditer est faible.

Utilisation de post-éditions pour corriger les sorties du système de traduction

A la lumière des résultats précedemment obtenus, nous avons tenté de fournir une meilleure compréhension de l'utilité des systèmes de post-édition statistique pour améliorer les résultats d'un système de traduction probabiliste. Pour cela, nous avons conduit une série d'expériences faisant varier différents paramètres en jeux dans l'application de tels systèmes.

En premier lieu, nous avons montré qu'un système de post-édition statistique appris sur des données de domaine général de taille moyenne ($\approx$ 9 000 phrases) n'apporte aucun gain en terme de score BLEU et TER quand il est appliqué sur les résultats d'un système de traduction probabiliste général. Avec de tels paramètres expérimen-

taux, l'utilisation d'hypothèses de traduction corrigées manuellement (configuration « réelle ») à la place de traduction professionnelles indépendantes du système (configuration « simulée ») donne des résultats légèrement plus satisfaisants.

Nous avons ensuite observé que l'ajout de données pour l'apprentissage du SPES n'est pas suffisant pour améliorer significativement les performances du système de traduction de référence. Quelle que soit la taille des données disponibles pour l'apprentissage sur SPES, il semble difficile d'améliorer et de corriger les résultats d'un système de traduction probabiliste généraliste à l'aide d'une post-édition statistique.

Si l'on compare le SPES appris sur un domaine général à un SPES appris sur un domaine spécifique, nous constatons que ce dernier nous permet d'atteindre des performances significativement meilleures. La post-édition statistique voit donc son efficacité croitre avec la spécificité du domaine sur lequel le système est appris. Dans ce contexte, le système peut donc être utilisé avec succès pour adapter un système de traduction probabiliste généraliste à un domaine spécifique. Nous avons néanmoins remarqué que ce gain de qualité était dû, en grande partie, au taux important de mots communs hors-vocabulaire du système général, reconnus et corrigés par le SPES spéficique au domaine.

Dans nos dernières expérimentations, nous comparons l'adaptation au domaine faite par le système de post-édition statistique à d'autres méthodes d'adaptation au domaine qui consistent à apprendre des modèles de traduction et des modèles de langage spécifiques au domaine puis à les interpoler avec les modèles généraux. Dans notre protocole, les méthodes d'adaptation utilisées partagent le même système de traduction de référence et les même corpus d'apprentissage et de test. La qualité des résultats des approches basées sur l'adaptation des modèles de langage et de traduction dépasse celle des résultats obtenus avec le SPES spécifique. Il est toutefois important de noter que ces méthodes d'adaptation au domaine nécessitent l'apprentissage d'un nouveau système de traduction probabiliste. On peut facilement imaginer des situations pratiques où il serait impossible de créer un nouveau système (c'est le cas d'un système de traduction utilisé comme une « boîte noire ») ou des situations où il serait plus avantageux de conserver un système de traduction généraliste sur lequel seraient appliqués plusieurs SPES adaptés à différents domaines.

Utilisation de post-éditions pour l'estimation de confiance en traduction automatique

Il est usuellement admis que moins une hypothèse de traduction automatique nécessite d'opérations de post-édition, plus sa qualité est estimée comme grande. A partir de ce constat, nous avons utilisé notre corpus de post-éditions pour apprendre des modèles d'estimation de confiance pour la traduction automatique. Par le biais de nos expérimentations, nous avons proposé des modèles capables de prédire un score de qualité d'une traduction fournie par un système automatique directement en fonction de l'effort ou du taux de post-édition nécessaire pour la transformer en une traduction répondant à nos critères de qualité. Nous avons estimé le taux de post-édition par la métrique TER calculée entre une hypothèse de traduction automatique et sa post-édition.

Nous avons extrait un ensemble de 18 fonctions indicatrices de qualité de notre corpus de 10 881 phrases et nous avons utilisé les post-éditions collectées pour apprendre deux types de Séparateurs à Vaste Marge (SVM) pour apprendre, d'une part, des modèles de régression destinés à prédire la valeur numérique du score TER associé à un couple *(phrase source, hypothèse de traduction)* donné, et d'autre part, des classifieurs destinés à associer une classe de qualité à ce même couple.

Les résultats obtenus montrent que les deux types d'estimateurs créés sur ces données produisent des modèles de qualité tant pour prédire le taux de post-édition nécessaire à une hypothèse de traduction donnée que pour lui affecter une classe de qualité à 5 niveaux. En effet, le meilleur résultat des systèmes de prédiction du taux de post-édition montrent un score MAE de 0,13, ce qui signifie que le modèle d'estimation est capable de prédire un score TER avec, en moyenne, un écart de 0,13 points comparé au score TER de référence calculé entre l'hypothèse de traduction et sa post-édition. Pour les systèmes de classification de la qualité de traduction, le taux de précision de plus de 48 % signifie que, près d'une fois sur deux, la classe de TER prédite par le système d'estimation de confiance est identique à la classe donnée pour référence dans le corpus. Ce résultat est renforcé par les scores MAE et RMSE obtenus (respectivement de 0,63 et 0,93) qui montrent que le modèle de prédiction de qualité attribue des classes qui sont, en moyenne, très proches (dans le sens ordinal) des classes de référence.

Les résultats de nos travaux montrent donc que le corpus de post-éditions collecté par nos soins, constitue un ensemble de données intéressant pour l'apprentissage de mesures de confiance pour la traduction automatique statistique.

Perspectives de recherche

Nos travaux ont montré que l'intégration de corrections humaines de traduction automatiques dans les systèmes représentait une problématique prometteuse mais aussi complexe. L'analyse des résultats obtenus nous a permis de détecter les apports ainsi que les limites des différentes techniques proposées dans nos travaux. Nous discutons ici des extensions possibles de nos travaux ou d'autres pistes de recherches liées à la problématique.

Vers l'analyse et la classifications d'erreurs

Une analyse approfondie du corpus de post-éditions collectées, et en particulier l'identification et le recensement des erreurs corrigées par la post-édition, pourrait permettre, en premier lieu, de détecter les erreurs de traduction du système et leurs origines, voire de procéder à une analyse fine des lacunes du système de traduction.

Au delà des techniques que nous avons présentées dans ce manuscrit, cette étude pourrait déboucher sur l'élaboration d'un système de classification des erreurs de traduction, à l'instar de ce qui est fait dans [Popovic et Ney 2007, Zhou et al. 2008, Popovic et Burchardt 2011]. Ces travaux s'inspirent du schéma de classification d'erreurs proposé dans [Vilar et al. 2006] pour élaborer un système qui identifie les erreurs de traduction et leur assigne une classe prédéfinie (parmi les suivantes : erreur d'inflection, erreur d'ordonnancement, omission, ajout et choix lexical incorrect) en utilisant

le résultat des métriques WER et PER.

Pour aller encore plus loin, l'idée serait d'utiliser ces classes typologiques pour détecter les erreurs présentes dans les résultats de traduction et de limiter leur propagation et/ou répandre leurs corrections sur d'autres corpus à l'aide d'un système correctif se basant sur la capture de schémas d'erreurs (« *patterns* »).

Il est important de noter que l'analyse d'erreurs étant intimement liée à la problématique de l'évaluation « locale » de la qualité d'une traduction, il est essentiel de concentrer une partie des efforts de recherche sur cette problématique. L'évaluation « locale » désigne l'évaluation au niveau d'une phrase ou d'une unité de traduction, par opposition à l'évaluation « globale » usuelle des systèmes qui se fait au niveau d'un corpus.

Vers un apprentissage actif et interactif

Au delà de l'intérêt évident du corpus de post-éditions que nous avons collecté pour étudier et analyser les corrections et les erreurs produites par le système de traduction automatique, celui-ci constitue une ressource précieuse pour simuler le résultat d'une annotation humaine dans les problématiques de reherche basées sur la post-édition de traductions automatiques.

C'est le cas par exemple de l'apprentissage actif, dont le principe est de réduire la quantité de données d'apprentissage annotées requise pour qu'un système de traduction automatique acquiert un certain niveau en terme de qualité de traduction [Haffari et Sarkar 2009, Haffari et al. 2009, Callison-Burch 2003]. Le but est alors de minimiser le coût de l'annotation en ne faisant appel à l'expertise humaine que lorsque l'utilité est élevée.

Par opposition à l'approche standard où la sélection des exemples d'apprentissage du système est faite de façon séquentielle ou aléatoire, dans l'approche active, le système automatique sélectionne les exemples d'apprentissage qui lui semblent les plus instructifs, c'est-à-dire les phrases les plus à même d'aider à ré-entrainer le système afin qu'il atteigne, rapidement, une certaine qualité. Les nouveaux exemples d'apprentissages ainsi sélectionnés sont annotés par des experts humains et ajoutés aux données d'apprentissage du système. Selon la satisfaction d'un critère d'arrêt, le processus est ré-itéré ou non.

De part leur lien étroit avec l'annotation humaine de données d'apprentissage et l'évaluation « locale » des traductions, les travaux décrits dans ce manuscrit peuvent directement s'inscrire dans le contexte de l'apprentissage actif et plus particulièrement dans l'étude des stratégies de sélection des exemples à annoter et des méthodes d'intégration des nouvelles annotations dans le système.

Pour conclure ce manuscrit, nous souhaiterons rappeler que la qualité des traductions produits par les systèmes automatiques n'est souvent pas suffisante pour permettre une automatisation complète du processus traductif, notamment à cause des erreurs liées à l'ambiguïté des mots, la prise en compte du contexte et la couverture du vocabulaire. L'intervention humaine reste souvent nécessaire pour corriger les erreurs commises par les systèmes, comme l'exprime Yves Champollion en 2001 : "[...] *since translation without understanding is meaningless, the future of the human translator is proof-sensing what a machine has pre-translated.*". [49]

Si l'on considère ce dernier fait, force est de constater que la post-edition manuelle des hypothèses de traduction du système automatique constitue, au jour d'aujourd'hui, une étape essentielle du processus de traduction (qu'il soit grand public ou professionnel). De ce fait, la ré-utilisation de ces corrections humaines de traductions automatiques pour améliorer les systèmes est, comme nous avons eu l'occasion de le constater, une problématique complexe mais néanmoins prometteuse qui représente un axe de recherche ouvert et actuel.

49. http ://www.bokorlang.com/journal/15mt.htm

Annexes

Annexe 1 Significativité des différences entre deux scores BLEU

La significativité statistique des différences entre les scores BLEU des expérimentations présentées dans ce manuscrit est évaluée selon la technique d'amorce par réechantillonage (ou *bootstrap resampling method* en anglais) proposée dans [Koehn 2004]. La méthode étant coûteuse à mettre en œuvre pour chacun des résultats obtenus, nous utilisons les résultats expérimentaux de l'étude de P. Koehn pour estimer le taux de confiance à accorder aux variations de scores BLEU. Le taux de confiance à accorder à la significativité statistique d'une différence (en pourcentage) entre deux scores BLEU, en tenant compte de la taille du corpus de test en nombre de phrases (on considère alors des phrases d'environ 30 mots en moyenne), est donnée par le tableau 1.1. Ainsi, par exemple, pour un corpus de test de 600 phrases, il sera nécessaire d'obtenir une différence de scores BLEU de 2,4 % pour que celle-ci soit jugée significative à 95 %, avec une confiance de 100 %.

Différence entre deux scores BLEU (en %)	Taille du corpus de test (en phrases)			
	100	300	600	3 000
5,1	97 %	100 %	100 %	100 %
4,3	85 %	100 %	100 %	100 %
2,7	65 %	97 %	100 %	100 %
2,4	53 %	91 %	100 %	100 %
2	31 %	65 %	96 %	100 %
1,6	33 %	60 %	84 %	100 %
1,5	24 %	48 %	74 %	100 %
0,5	7 %	12 %	10 %	30 %

Table 1.1 – Taux de confiance à accorder à la significativité statistique à 95 % d'une différence entre deux scores BLEU, en fonction de la taille du corpus de test, selon [Koehn 2004].

Annexe 2 Corpus post-édité et distances d'édition

Répartition des traductions du corpus post-édité selon leurs distances d'édition

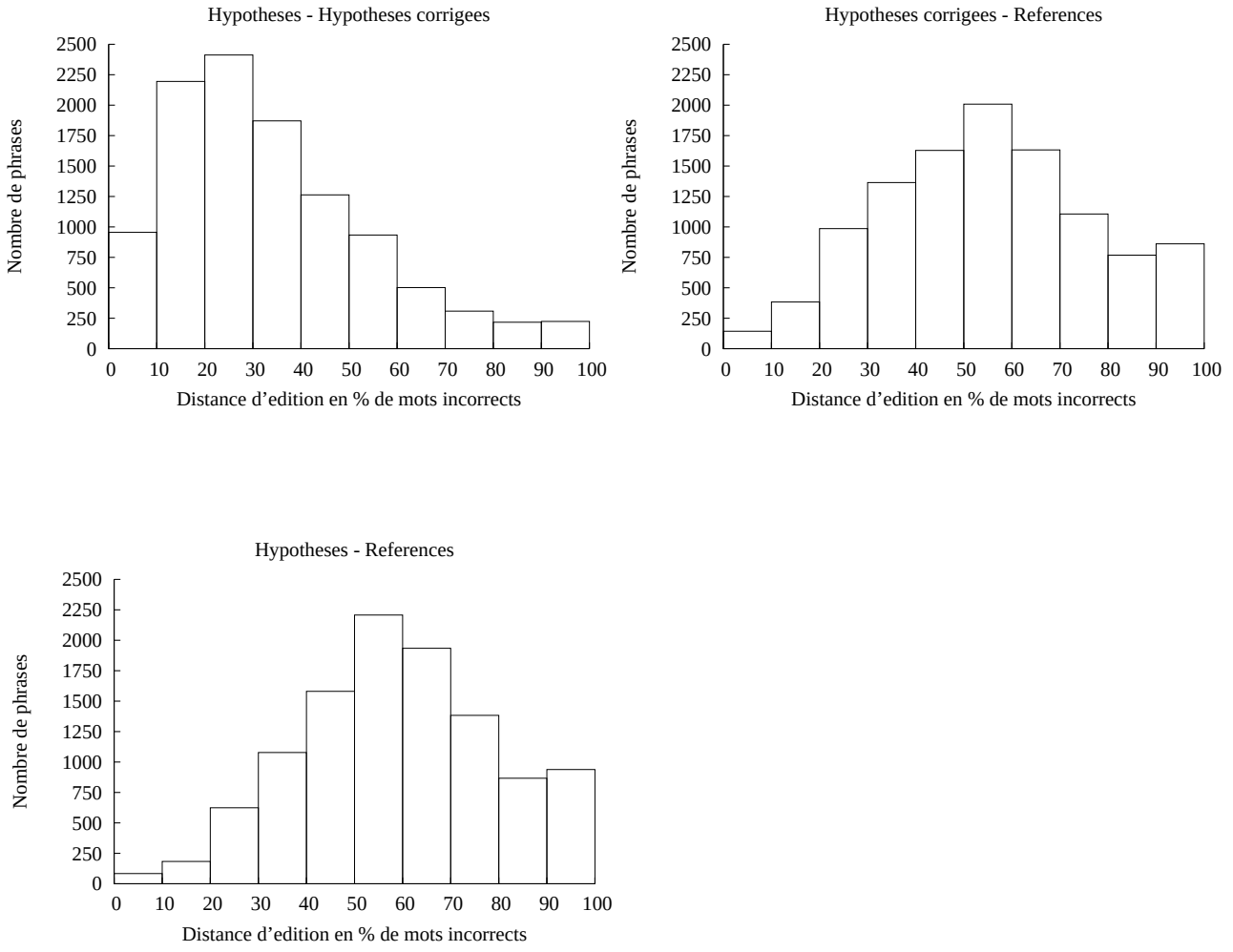


FIGURE 2.1 – Répartition des 10 881 phrases du corpus post-édité en fonction de la distance d'édition entre les différents types de traduction

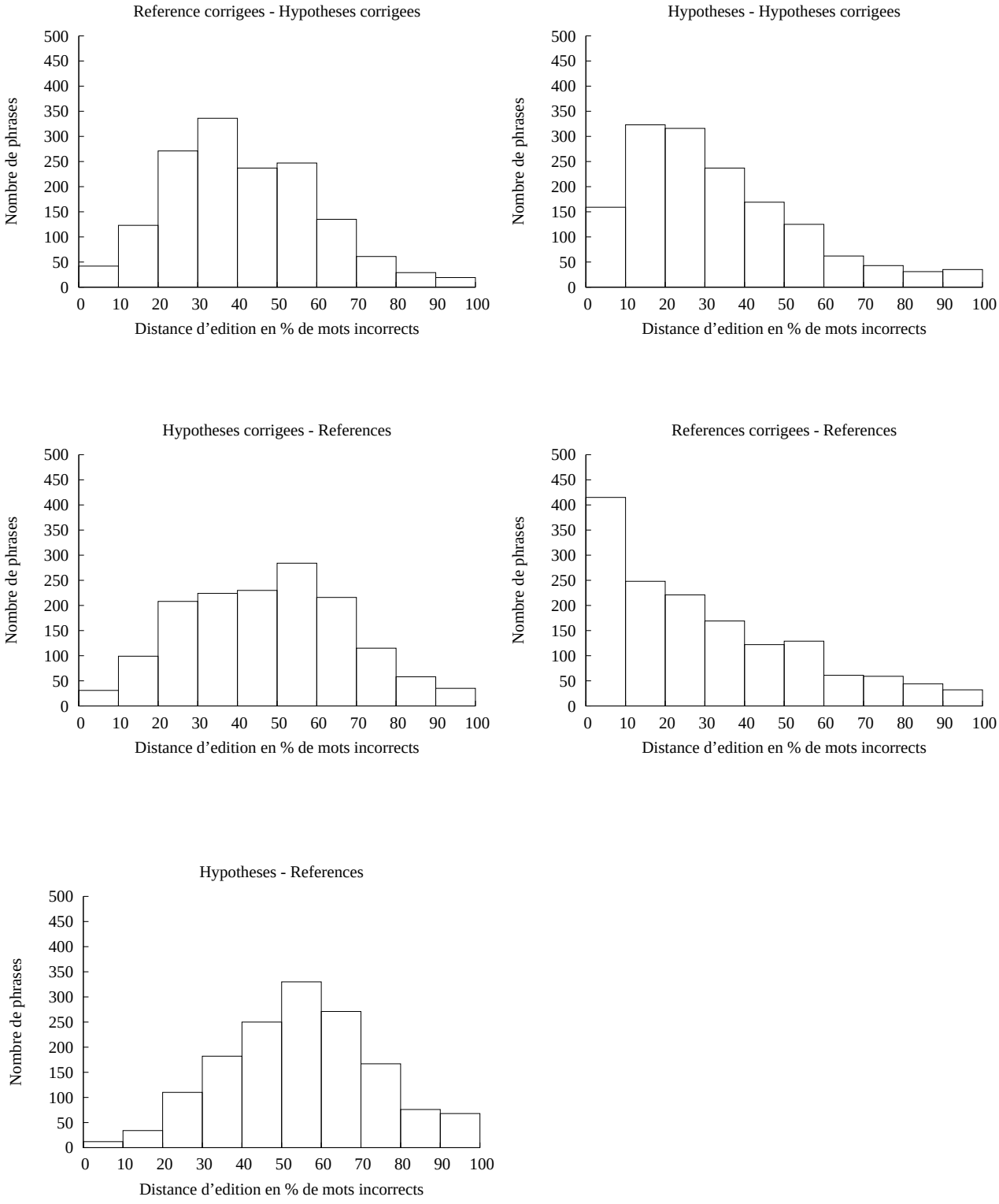


FIGURE 2.2 – Répartition des 1 500 phrases du corpus post-édité en fonction de la distance d'édition entre les différents types de traduction

Annexe 3 Exemples de phrases extraites de notre corpus de post-éditions

Cette annexe contient des exemples de phrases extraites de notre corpus de post-édition. Les tableaux sont décrits ci-après.

TAB 2.1 : phrases du corpus pour lesquelles la traduction professionnelle de référence fournie dans le corpus parallèle représente une traduction « éloignée »de la phrase source de part sa traduction non-littérale ;

TAB 2.2 : traductions dont la post-édition a été validée après plus de 4 tentatives de soumission (ici, la traduction de la phrase source peut être l'hypothèse de traduction générée par notre système de traduction de référence ou la traduction professionnelle de référence fournie dans le corpus parallèle) ;

TAB 2.3 : jugements humains des post-éditions collectées ;

TAB 2.4 : comparaison de post-édition professionnelles et non-professionnelles pour une même hypothèse de traduction. L'astérisque (*) désigne la post-édition jugée comme meilleure lors de notre évaluation. L'absence d'astérisque indique que les post-éditions professionnelles et non-professionnelles ont été jugées comme équivalentes.

Phrase source	Traduction de référence	Choix de traduction
• *Le prix du baril de pétrole baissera sans doute avec l'apparition d'une nouvelle source d'approvisionnement*, d'où une dépendance accrue des USA, notamment vis-à-vis du Moyen-Orient.	• **This, in turn, would mean an increase** in US dependence on imported oil, especially from the Middle East.	- "l'apparition [...] approvisionnement" ⇒ "this" - ajout : ∅ ⇒ "in turn" - modifications grammaticales et syntaxiques
• Malgré les apparences, ces objectifs sont toujours les mêmes *aujourd'hui*, sous l'administration Bush.	•Despite appearances, **none of** these objectives **has changed** under the Bush administration.	- forme active ⇒ forme passive - supression : "aujourd'hui" ⇒ ∅
• Beaucoup d'observateurs croient qu'avec l'invasion de l'Irak, Bush a choisi une autre voie.	• Many observers believe that Bush has set a new course because the invasion of Iraq **seems to fly in the face of these objectives**.	- ajout d'un syntagme verbal : "seems [...] objectives" - permutation syntaxique
• Il savait parfaitement ce qu'il devait à l'or noir.	• He **understood** perfectly **the role that** oil played **in his power**.	- "savait"⇒ "understood" - "ce qu'il devait à"⇒ "the role [...] his power"
• Bien entendu, les Américains contestent *les méthodes à utiliser*.	• Of course, Americans disagree on **how to do this**.	- modification de la catégorie grammaticale du syntagme : "les méthodes à utiliser"⇒ "how to do this"

TABLE 3.2 – Exemples de phrases du corpus pour lesquelles la traduction professionnelle de référence fournie dans le corpus parallèle représente une traduction « éloignée » et/ou non littérale de la phrase source.

Phrase source	Traduction	Traduction corrigée
• Chaque année, nous perdons 14,6 millions d'hectares de forêts, soit une zone presque aussi grande que quatre fois la Suisse.	• Every year, we lose 14.6 million hectares (56,000 square miles) of forests – an area almost four times the area of Switzerland.	• Every year, we lose 14.6 million hectares of forests, that is an area almost four times the size of Switzerland.
• Sans faire un grand match, les Jaunes et Noirs se sont créés le plus grand nombre d'occasions et ont réussi à en transformer une pour le gain du match.	• Without a great match, the yellow and black people have created the largest number of occasions and have managed to turn a profit for the match.	• Without playing a great match, the Yellows and the Blacks created for themselves the greater number of opportunities and succeeded in converting one into the match victory.
• En effet, chacune d'elles n'expliquait les phénomènes constatées lors de ventes aux enchères que par le plaisir de gagner et la parade devant le risque.	• Indeed, each of them explained the phenomena noted at auctions by the pleasure of winning and the parade before the risk.	• Indeed, each of them explained the phenomena noticed during auctions by the pleasure of winning and the parade before the risk.
• Or, il est également important si vous avez toujours sur vous des écouteurs ou bien un handsfree avec connexion.	• It is also important if you are always on the headphones or a handsfree with connection.	• Yet, it is also important if you always wear headphones or a handsfree with connection.

TABLE 3.3 – Exemples de traductions dont la post-édition a été validée après plus de 4 tentatives de soumission (ici, la traduction de la phrase source peut être l'hypothèse de traduction générée par notre système de traduction de référence ou la traduction professionnelle de référence fournie dans le corpus parallèle)

Phrase source	Hypothèse de traduction	Hypothèse de traduction corrigée	Caractère correctif	Caractère traductif
• Mais il est fort peu probable que ces deux tactiques permettent de développer le secteur de haute technologie européen.	• But it is highly unlikely that these tactics will develop the high-tech sector.	• But it is highly unlikely that these tactics will help develop the European high-tech sector.	• Correctrice	• Erronée
• Habituellement, le parti Républicain américain défendait des budgets équilibrés.	• Usually, America's Republican Party had balanced budgets.	•Usually, America's Republican Party called for balanced budgets.	•Correctrice	•Imprécise
• La durée de vie des batteries est également décisive.	• The life of batteries is also crucial.	• The long lasting of batteries is also crucial.	• Equivalente	• Imprécise
• Mais du point de vue américain, le rôle international du dollar était un piège.	• But from the American point of view, the international role of the dollar was a trap.	• But from the American point of view, the international role of the dollar was a trap.	• Equivalente	• Bonne
• Il est vain de s'accrocher à des illusions.	• There is no use holding onto illusions.	• There is no use in holding on to illusions.	• Correctrice	• Bonne
• Les prix des billets d'avion et des hôtels chutent.	• The airline ticket prices and hotels are dropping.	• The airline tickets prices and hotels are dropping	• Equivalente	• Erronée

TABLE 3.4 – Exemples de jugements humains des post-éditions collectées

Phrase source	Hypothèse de traduction et	Post-édition non-professionnelle	Post-édition professionnelle
• L'épopée de Meat Loaf devra être honorée.	• The saga of Meat Loaf will be honoured.	• * The saga of Meat Loaf must be honored.	• The saga of Meat Loaf will be honoured.
• Autrement, on risquerait que d'autres pays en subissent les incidences.	• Otherwise, we risk that other countries are suffering the effects.	• * Otherwise, we would risk that other countries suffer the effects.	• Otherwise, we risk that other countries will suffer the effects.
• Danseuse, actrice et dessinatrice, tu es une femme très polyvalente : n'est-ce pas ?	• Dancer, actress and dessinatrice, you are a woman very versatile : is it not ?	• * Dancer, actress and draftwoman, you are a very versatile woman : aren't you ?	• Dancer, actress and painter, you are a very versatile woman : is it not ?
• La première boutique se trouve à Londres.	• The first shop in London.	• * The first shop is situated in London.	• The largest is in London.
• Paul Newman (1925-2008) - Mort d'une icône engagée	• Paul Newman (1925-2008) - death of an icon committed	• * Paul Newman (1925-2008) - death of a committed icon	• Paul Newman (1925-2008) - death of an icon
• Il a eu une belle ovation.	• It was a nice ovation.	• He had a nice ovation.	• * He received a nice ovation.
• Le changement de stratégie est le bienvenue.	• The change of strategy is welcome.	• The strategy change is welcome.	• The change of strategy is welcome.
• Factures établies délibérément	• Bills set deliberately	• Invoices set deliberately	• Bills made deliberately
• C'est sûr qu'il y en a plus d'un qui parle pendant la nuit.	• It is sure that there are more than one who speaks during the night.	• It is sure that there is more than one who speaks during the night.	• It is certain that more than one speaks during the night.

TABLE 3.5 – Exemples de comparaison de post-éditions professionnelles et non-professionnelles pour une même hypothèse de traduction. L'astérisque (*) désigne la post-édition jugée comme meilleure lors de notre évaluation. L'absence d'astérisque indique que les post-éditions professionnelles et non-professionnelles ont été jugées comme équivalentes

Annexe 4 Evaluation de la suspicion de fraude avec *Google Translate*

Heuristique pour l'évaluation de la suspicion de fraude par traduction avec le logiciel en ligne *Google Translate*[50].

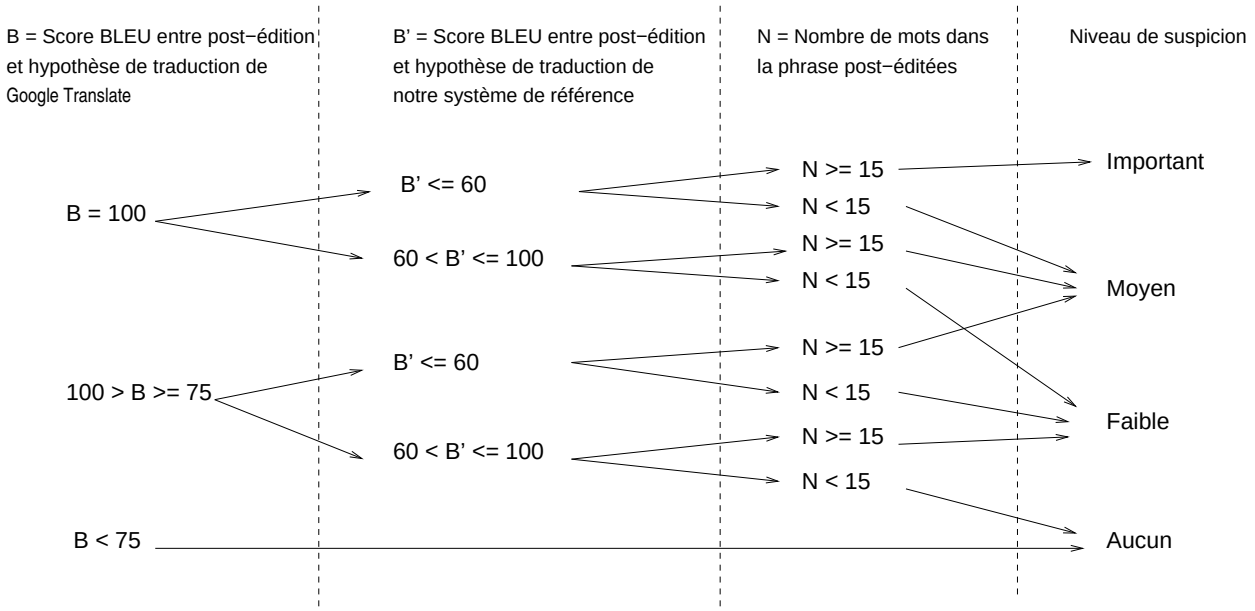


Figure 4.3 – Heuristique pour l'évaluation de la suspicion de fraude par traduction avec le logiciel en ligne *Google Translate*

50. http ://translate.google.fr

Annexe 5 Publications de l'auteure

Conférences internationales avec comité de lecture et publication des actes

Marion Potet, Laurent Besacier, Hervé Blanchon et Marwen Azouzi.(2012). *Toward a Better Understanding of Statistical Post-Edition Usefulness*. Dans les actes de la 9^{ime} conférence intitulée "International Workshop on Spoken Language Translation (IWSLT)". Hong-Kong, 6-7 Décembre 2012. 8 p.

Marion Potet, Emmanuelle Esperança-rodier, Laurent Besacier et Hervé Blanchon.(2012). *Collection of a Large Database of French-English SMT Output Corrections*. Dans les actes de la 8^{ime} conférence intitulée "International Conference on Language Resources and Evaluation (LREC)". Istanbul, Turquie, 23-25 Mai 2012. 6 p.

Marion Potet, Emmanuelle Esperança-rodier, Laurent Besacier et Hervé Blanchon.(2011). *Preliminary Experiments on Using Users' Post-Editions to Enhance a SMT System*. Dans les actes de la 15^{ime} conférence intitulée "European Association for Machine Translation (EAMT)". Louvain, Belgique, 05-07 Mai 2011. Vol. 1/1 : pp. 162-168.

Campagnes d'évaluation internationales avec comité de lecture et publication des actes

Marion Potet, Raphaël Rubino, Benjamin Lecouteux, Stéphane Huet, Laurent Besacier, Hervé Blanchon et Fabrice Lefèvre.(2011). *The LIGA (LIG/LIA) Machine Translation System for WMT 2011*. Dans les actes de la conférence intitulée "EMNLP, 6^{ime} Workshop on Statistical Machine Translation (WMT)". Edimbourg, Royaume-Uni. 27-31 Juillet 2011. Vol. 1/1 : pp. 440-446.

Laurent Besacier, Haithem Afli, Do Thi Ngoc Diep , Hervé Blanchon and **Marion Potet**.(2010). *LIG Statistical Machine Translation Systems for IWSLT 2010*. Dans les actes de la 7^{ime} conférence intitulée "International Workshop on Spoken Language Translation(IWSLT)". Paris, France. 02-03 Décembre 2010. Vol. 1/1 : pp. -.

Potet Marion, Laurent Besacier et Hervé Blanchon (2010). *The LIG machine translation system for WMT 2010*. Dans les actes de la conférence intitulée "ACL, 5^{ime} Workshop on statistical Machine Translation (WMT)". Uppsala, Suède. 11-17 juillet 2010. Vol. 1/1 : pp. 167-172.

Conférences nationales avec comité de lecture et publication des actes

Potet Marion. *Méta-moteur de traduction automatique : proposition d'une métrique pour le classement de traductions.*(2009). Dans les actes de la 14^{ime} Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Senlis, France. 23-26 juin 2009. Vol. 1/1 : pp. 373-382.

Rapports

Marion Potet. *Optimisation pour les systèmes de traduction automatiques probabilistes par adaptation dynamique des paramètres du modèle log-linéaire.*(2009). Rapport de fin d'étude de Master 2 Recherche, Université Joseph-Fourier, Grenoble, France. 65 pages. Juin 2009.

Marion Potet. *Proposition d'une métrique pour le classement de traductions et création d'un méta-moteur de traduction automatique.*(2008). Rapport de fin d'étude de Master 2 Professionel, Université Pierre Mendès-France, Grenoble, France. 60 pages. Août 2008.

Bibliographie

[Albrecht et Hwa 2007] Joshua Albrecht et R. Hwa (2007). Regression for Sentence-Level MT Evaluation with Pseudo References. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 296–303, Prague, Czech Republic.

[Allen et Hogan 2000] Jeffrey Allen et C. Hogan (2000). Toward the Development of a Post-Editing Module for Machine Translation raw Output : a Controlled Language Perspective. *Proceedings of the 3rd International Controlled Language Applications Workshop (CLAW'00)*, pages 62–71, Washington DC, USA.

[Babych et Hartley 2004] Bogdan Babych et A. Hartley (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 621–628, Barcelona, Spain.

[Banerjee et Lavie 2005] S. Banerjee et A. Lavie (2005). METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the international Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL'05)*, pages 65–72, Ann Arbor, Michigan, USA.

[Bar-Hillel 1960] Yehoshua Bar-Hillel (1960). The Present Status of Automatic Translation of Languages. Advances in Computers, 1 :91–163.

[Barrachina et al. 2009] Sergio Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, E. Vidal et J.-M. Vilar (2009). Statistical Approaches to Computer-Assisted Translation. Computational Linguistics, 35(1) :3–28.

[Béchara et al. 2011] Hanna Béchara, Y. Ma et J. Van Genabith (2011). Statistical Post-Editing for a Statistical MT System. *Proceedings of the 13th Machine Translation Summit (MT SUMMIT XIII)*, pages 308–315, Xiamen, China.

[Bertoldi et al. 2009] Nicolas Bertoldi, B. Haddow et J.-B. Fouet (2009). Improved Minimum Rate Training in Moses. The Prague Bulletin of Mathematical Linguistics, 91 :7–16.

[Blanchon et Boitet 2007] Hervé Blanchon et C. Boitet (2007). Pour l'évaluation externe des systèmes de TA par des méthodes externes fondées sur la tâche. Traitement Automatique des Langues, 48(1) :175–204.

[Blanchon et al. 2009] Hervé Blanchon, C. Boitet et C.-P. Huynh (2009). A Web Service Enabling Gradable Post-edition of Pre-translations Produced by Existing Translation Tools : Practical Use to Provide High Quality Translation of an Online Encyclopedia. *Proceedings of the International Association for Machine Translation hosted by the Association for Machine Translation in the America (MT Summit XII 2009)*, pages 20–27, Ottawa, Canada.

[Blatz et al. 2003] John Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis et N. Ueffing (2003). Confidence Estimation for Machine Translation. Technical report, CLSP Summer Workshop, Johns Hopkins University.

[Brown et al. 1990] Peter Brown, J. Coke, S. D. Pietra, V. D. Pietra, F. Jelinek, J. Lafferty, R. Mercer et P. Roosin (1990). A Statistical Approach to Machine Translation. *Proceedings of the 28st Annual Meeting of the Association for Computational Linguistics on Human Language Technology (ACL-HLT'90)*, volume 16, pages 79–85, Pittsburgh, Pennsylvania, USA.

[Brown et al. 1991] Peter Brown, J. Coke, S. D. Pietra, V. D. Pietra et R. Mercer (1991). A Statistical Approach to Sense Disambiguation in Machine Translation. *Proceedings of the 29st Annual Meeting of the Association for Computational Linguistics on Human Language Technology (ACL-HLT'91)*, pages 146–151, Berkeley, California, USA.

[Brown et al. 1993] Peter Brown, J. Coke, S. D. Pietra, V. D. Pietra et R. Mercer (1993). The Mathematics of Statistical Machine Translation : Parameter Estimation. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics on Human Language Technology (ACL-HLT'93)*, volume 19, pages 263–311, Ohio, USA.

[Callison-Burch 2003] Chris Callison-Burch (2003). *Active learning for statistical machine translation*. Thèse, Edinburgh University.

[Callison-Burch et al. 2007] Chris Callison-Burch, C. Fordyce, P. Koehn, C. Monz et J. Schroeder (2007). (Meta-)Evaluation of Machine Translation. *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

[Callison-Burch et al. 2008] Chris Callison-Burch, C. Fordyce, P. Koehn, C. Monz et J. Schroeder (2008). Further Meta-Evaluation of Machine Translation. *Proceedings of the 3rd Workshop on Statistical Machine Translation (StatMT'08)*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

[Callison-Burch et al. 2010] Chris Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki et O. Zaidan (2010). Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. *Proceedings of the 5th Workshop on Statistical Machine Translation (WMT'10)*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

[Callison-Burch et al. 2012] Chris Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut et L. Specia (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the 7th Workshop on Statistical Machine*

Translation (WMT'12), pages 10–51, Montréal, Canada. Association for Computational Linguistics.

[Callison-Burch et al. 2009] Chris Callison-Burch, P. Koehn, C. Monz et J. Schroeder (2009). Findings of the 2009 Workshop on Statistical Machine Translation. *Proceedings of the 4th Workshop on Statistical Machine Translation (StatMT'09)*, pages 1–28, Athens, Greece.

[Callison-Burch et al. 2006] Chris Callison-Burch, M. Osborne et P. Koehn (2006). Re-evaluating the Role of Bleu in Machine Translation Research. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 249–256, Budapest, Hungary.

[Cameron 2007] S. Fordyce Cameron (2007). Overview of the IWSLT 2007 Evaluation Campaign. *Proceedings of the 4th International Workshop on Spoken Language Technology : workshop opening (IWSLT'07)*, Trento, Italy.

[Cettolo et Federico 2004] M. Cettolo et M. Federico (2004). Minimum Error Training of Log-Linear Translation Models. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'04)*, pages 103–106, Kyoto, Japan.

[Chang et Lin 2011] Chih-Chung Chang et C.-J. Lin (2011). LIBSVM : A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, 2(3) :27 :1–27 :27.

[Chen et Goodman 1999] Stanley F. Chen et J. Goodman (1999). An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4) :359–393.

[Chiang 2007] David Chiang (2007). Hierarchical Phrase-Based Translation. Computational Linguistics, 33(2) :201–228.

[Chiang 2012] David Chiang (2012). Hope and Fear for Discriminative Training of Statistical Translation Models. Machine Learning Research, 13(1) :1159–1187.

[Costa-jussà et Fonollosa 2006] Marta R. Costa-jussà et J. A. R. Fonollosa (2006). Statistical Machine Reordering. *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 70–76, Sydney, Australia.

[Dempster et al. 1977] A.P. Dempster, N. Laird et D. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological)., 39(1) :1–38.

[Doddington 2002] George Doddington (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proceedings of the 2nd international conference on Human Language Technology Research*, pages 138–145, San Diego, California, USA.

[Dugast et al. 2007] Loic Dugast, J. Senellard et P. Koehn (2007). Statistical Post-Editing on Systran's Rule-Based Translation System. *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*, pages 220–223, Prague, Czech Republic.

[Dugast et al. 2009] Loic Dugast, J. Senellart et P. Koehn (2009). Statistical Post-Editing and Dictionary Extraction : Systran/Edinburg Submissions for WMT2009. *Proceedings of the 4th Workshop on Statistical Machine Translation (ACL-WMT'09)*, pages 110–114, Athens, Greece.

[Déchelotte 2010] Daniel Déchelotte (2010). Analysis of Translation Suggestions on Reverso Translation Engines : Initial Findings. FAUST project report.

[Elming 2006] Jakob Elming (2006). Transformation-Based Corrections of Rule-Based MT. *Proceedings of the 11th Conference of the European Association on Machine Translation (EAMT'06)*, pages 219–226, Oslo, Norway.

[Fort et al. 2011] Karën Fort, G. Adda et K. B. Cohen (2011). Amazon Mechanical Turk : Gold Mine or Coal Mine ? Journal of Computational Linguistics, 37 :413–420.

[Foster et Kuhn 2009] George Foster et R. Kuhn (2009). Stabilizing Minimum Error Rate Training. *Proceedings of the 4th Workshop on Statistical Machine Translation (StatMT'09)*, pages 242–249, Athens, Greece.

[Gandrabur et Foster 2003] Simona Gandrabur et G. Foster (2003). Confidence Estimation for Translation Prediction. *Proceedings of the 7th Conference on Natural Language Learning (CONLL) at HLT-NAACL 2003 - Volume 4* , pages 95–102, Edmonton, Canada.

[Garcia 2011] Ignacio Garcia (2011). Translating by Post-Editing : is it the Way Forward ? Journal of Machine Translation, 25(3) :217–237.

[Gelas et al. 2011] Hadrien Gelas, S. T. Abate, L. Besacier et F. Pellegrino (2011). Evaluation of Crowdsourcing Transcriptions for African Languages. *Proceedings of the conference on Human Language Technologies for Development (HLTD'11)*, pages 128–133, Alexandrie, Egypt.

[Guzmàn 2007] Raphael Guzmàn (2007). Automating MT post-editing using regular expressions. Multilingual, 18(6) :49–52.

[Haffari et al. 2009] Gholamreza Haffari, M. Roy et A. Sarkar (2009). Active Learning for Statistical Phrase-Based Machine Translation. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*, pages 415–423, Boulder, Colorado.

[Haffari et Sarkar 2009] Gholamreza Haffari et A. Sarkar (2009). Active Learning for Multilingual Statistical Machine Translation. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 181–189, Suntec, Singapore.

[Hasler et al. 2011] Eva Hasler, B. Haddow et P. Koehn (2011). Margin Infused Relaxed Algorithm for Moses. The Prague Bulletin of Mathematical Linguistics, 96(1) :69–78.

[He et al. 2010] Yifan He, Y. Ma, J. Roturier, A. Way et J. van Genabith (2010). Improving the Post-Editing Experience using Translation Recommendation :

a User Study. *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA'10)*, Denver, CO, USA.

[Huynh et al. 2008] Cong-Phap Huynh, C. Boitet et H. Blanchon (2008). SECTra_w.1 : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco.

[Diaz de Ilarraza et al. 2008] Arantza Diaz de Ilarraza, G. Labaka et K. Sarasol (2008). Statistical Post-Editing : a Valuable Method in Domain Adaptation of RBMT Systems for Less-Resourced Languages. *Proceedings of international conference on Mixing Approaches to Machine Translation (MATMT'08)*, pages 35–40, Donostia-San Sebastian, Spain.

[Isabelle et al. 2007] Pierre Isabelle, C. Goutte et M. Simard (2007). Domain Adaptation of MT Systems through Automatic Post-Editing. *Proceedings of the 11th Machine Translation Summit (MT Summit XI)*, pages 255–261, Copenhagen, Denmark.

[Johnson et al. 2007] H. Johnson, J. Martin, G. Foster et R. Kuhn (2007). Improving Translation Quality by Discarding Most of the Phrasetable. *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 967–975, Prague, République Tchèque.

[Kay 1973] Martin Kay (1973). The MIND system. *Courant Computer Science Symposium 8 : Natural Language Processing*, pages 155–188, Algorithmics Press, New York.

[Knight et Chander 1994] Kevin Knight et I. Chander (1994). Automated Postediting of documents. *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI '94)*, pages 779–784, Seattle, USA.

[Koehn 2005a] Philip Koehn (2005a). Europarl : A Parallel Corpus for Statistical Machine Translation. *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

[Koehn 2004] Philipp Koehn (2004). Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the international Conference on empirical methods in natural language processing (EMNLP'04)*, pages 388–395, Barcelona, Spain.

[Koehn 2005b] Philipp Koehn (2005b). Europarl : A Parallel Corpus for Statistical Machine Translation. *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

[Koehn 2009] Philipp Koehn (2009). A Web-Based Interactive Computer Aided Translation Tool. *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore. Association for Computational Linguistics.

[Koehn 2011] Philipp Koehn (2011). What is a Better Translation ? Reflections on Six Years of Running Evaluation Campaigns. Communcation orale dans le premier colloque Tralogy (Tralogy I - Session 5). Auditorium du CNRS, Paris.

[Koehn et al. 2007] Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C.

Dyer, O. Bojar, A. Constantin et E. Herbst (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45st Annual Meeting of the Association for Computational Linguistics on Human Language Technology (ACL-HLT'07)*, pages 177–180, Prague, Czech Republic.

[Koehn et al. 2003] Philipp Koehn, F. J. Och, et D. Marcu (2003). Statistical Phrase-Based Translation. *Proceedings of the 41st Annual Meeting of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL'03)*, volume 1, pages 48–54, Edmonton, Canada.

[Koehn et Schroeder 2007] Philipp Koehn et J. Schroeder (2007). Experiments in Domain Adaptation for Statistical Machine Translation. *Proceedings of the 2nd Workshop on Statistical Machine Translation, Association for Computational Linguistics (ACL-StatMT '07)*, pages 224–227, Prague, Czech Republic.

[Koponen 2012] Maarit Koponen (2012). Comparing Human Perceptions of Post-editing Effort with Post-editing Operations. *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 181–190, Montréal, Canada. Association for Computational Linguistics.

[Kuhn et al. 2010] Roland Kuhn, P. Isabelle, C. Goutte, J. Senellart, M. Simard et N. Ueffing (2010). Recent Advances in Automatic Post-Editing. Journal of Multilingual computing and technology, 21(1) :43–46.

[Kulesza et Shieber 2004] Alex Kulesza et S. M. Shieber (2004). A Learning Approach to Improving Sentence-Level MT Evaluation. *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'04)*, pages 75–84, Baltimore, MD, US.

[Lagarda et al. 2009] Alabau Lagarda, S. Casacuberta et E. Diaz-de Liano (2009). Statistical Post-Editing of a Rule-based Machine Translation System. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09)*, pages 217–220, Boulder, Colorado, USA.

[Langlais et al. 2000] Philippe Langlais, G. Foster et G. Lapalme (2000). Transtype : a Computer-Aided Translation Typing System. *Proceedings of the NAACL-ANLP Workshop on Embedded machine translation - Volume 5*, pages 46–51, Seattle, Washington, USA.

[Lavie et al. 2004] Alon Lavie, K. Sagae et S. Jayaraman (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. *Proceedings of the 6th conference of the Association for Machine Translation in America Conference (AMTA'04)*, pages 134–143, Washington, USA.

[Lin et Och 2004] Chin-Yew Lin et F. J. Och (2004). Orange : a Method for Evaluating Automtic Evaluation Metrics for Machine Translation. *Proceedings of the 20th international conference on Computational Linguistics (COLING'04)*, pages 740–747, Stroudsburg, PA, USA.

[Liu et Gildea 2005] Ding Liu et D. Gildea (2005). Syntactic Features for Evaluation of Machine Translation. *Proceedings of the association for Computational Linguitics,*

Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 25–32, Ann Arbor, Michigan, USA.

[Liu et Gildea 2006] Ding Liu et D. Gildea (2006). Stochastic Iterative Alignement for Machine Translation Evaluation. *Proceedings of the International Conference on Computational Linguistics/Association for Computational Linguistics (COLING-ACL'06)*, pages 539–546, Sydney, Australia.

[Lock et Booth 1955] William N. Lock et A. D. Booth (1955). *Machine translation of languages : fourteen essays*. Technology Press.

[Luong 2012] Ngoc Quang Luong (2012). Integrating lexical, syntactic and system-based features to improve Word Confidence Estimation in SMT. *Proceedings of the 14th national French conference untitled « Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues » (RECITAL'12)*, pages 43–56, Grenoble, France. ATALA/AFCP.

[Mauser et al. 2006] Arne Mauser, R. Zens, E. Matusov, S. Hasan et H. Ney (2006). The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. *Proceedings of the 3th International Workshop on Spoken Language Translation (IWSLT'06)*, pages 103–110, Kyoto, Japan.

[Melby 1981] A. K. Melby (1981). Translators and Machines - Can they cooperate ? META, 26(1) :23–34.

[Mohit et Hwa 2007] Behrang Mohit et R. Hwa (2007). Localization of Difficult-to-translate Phrases. *Proceedings of the 2nd Workshop on Statistical Machine Translation (StatMT'07)*, pages 248–255, Prague, Czech Republic.

[Moore et al. 2008] Moore, C. Robert et C. Quirk (2008). Random Restarts in Minimum Error Rate Training for Statistical Machine Translation. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, pages 585–592, Manchester, UK.

[Nießen et al. 2000] S. Niessen, F. Och, G. Leusch et H. Ney (2000). An Evaluation Tool for Machine Translation : Fast Evaluation for MT Search. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, pages 39–45, Athènes, Grèce.

[Och 2003] Franz Josef Och (2003). Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics on Human Language Technology (ACL-HLT'03) - Volume 1*, pages 160–167, Sapporo, Japan.

[Och et al. 2004] Franz Josef Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin et D. Radev (2004). A Smorgasbord of Features for Statistical Machine Translation. *Proceedings of the Meeting of the North American chapter of the Association for Computational Linguistics and the Human Language Technology Conference (HLT-NAACL'04)*, pages 161–168, Boston, Massachusetts.

[Och et Ney 2002] Franz Josef Och et H. Ney (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL'02), pages 295–302, Philadelphia, Pennsylvania.

[Och et Ney 2003] Franz Josef Och et H. Ney (2003). A Systematic Comparison of Various Statistical Alignment Models. Journal of Computational Linguistics, 29(1) :19–51.

[Och et al. 1999] Franz Josef Och, C. Tillmann et H. Ney (1999). Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, USA.

[Oepen et al. 2007] Stephan Oepen, E. Velldal, J. T. Lønning, P. Meurer, V. Rosén et D. Flickinger (2007). Towards Hybrid Quality-Oriented Machine Translation. *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*, pages 144–153, Skovde, Sweden.

[Olsson 2009] Frederik Olsson (2009). A Literature Survey of Active Machine Learning in the Context of Natural Language Processing. Technical report, SICS : Swedish Institute of Computer Science, Sweden.

[Papineni et al. 2002] Kishore Papineni, S. Roukos, T. Ward et W.-J. Zhu (2002). BLEU : A method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics on Human Language, Technology session : Machine translation and evaluation (ACL-HLT'02)*, pages 311–318, Philadelphia, Pennsylvania, USA.

[Pierce et al. 1966] John R. Pierce, J. B. Carroll, E. P. Hamp, D. G. Hays, C. F. Hockett, A. G. Oettinger et A. Perlis (1966). Language and Machines : Computers in Translation and Linguistics. Technical report Publication 1416, The Automatic Language Processing Advisory Committee (ALPAC). Division of Behavioural Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

[Pighin et al. 2012a] Daniele Pighin, L. Màrquez et L. Formiga (2012a). The FAUST Corpus of Adequacy Assessments for Real-World Machine Translation Output. *Proceedings of the 8th international Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

[Pighin et al. 2012b] Daniele Pighin, L. Màrquez et J. May (2012b). An Analysis (and an Annotated Corpus) of User Responses to Machine Translation Output. *Proceedings of the 8th international Conference on Language Resources and Evaluation(LREC'12)*, Istanbul, Turkey.

[Popovic et Burchardt 2011] Maja Popovic et A. Burchardt (2011). From Human to Automatic Error Classification for Machine Translation Output. *proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT'11)*, pages 38–44, Leuven, Belgium.

[Popovic et Ney 2007] Maja Popovic et H. Ney (2007). Word Error Rates : Decomposition over POS classes and Applications for Error Analysis. *proceedings of*

the 2nd ACL Workshop on Statistical Machine Translation (WMT'07), pages 48–55, Prague, Czech Republic.

[Potet et al. 2010] Marion Potet, L. Besacier et H. Blanchon (2010). The LIG Machine Translation System for WMT 2010. *Proceedings of the joint 5th Workshop on Statistical Machine Translation and Metrics MATR, ACL Workshop (ACL-WMT'10)*, pages 11–17, Uppsala, Suède.

[Potet et al. 2012] Marion Potet, E. Esperanca-Rodier, L. Besacier et H. Blanchon (2012). Collection of a Large Database of French-English SMT Output Corrections. *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey.

[Quirk 2004] Christopher B. Quirk (2004). Training a Sentence-Level Machine Translation Confidence Metric. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 825–828, Lisbon, Portugal.

[Rafalovitch et Dale 2009] A. Rafalovitch et R. Dale (2009). United Nations General Assembly Resolutions : A six-langage Parallel Corpus. *Proceedings of the 12th Machine Translation Summit (MT Summit XII)*, pages 292–299, Ottawa, Canada.

[Raybaud et al. 2011] Sylvain Raybaud, D. Langlois et K. Smaïli (2011). "This sentence is wrong." Detecting errors in machine-translated sentences. Machine Translation, 25(1) :1–34.

[Rubino et al. 2012] Raphaël Rubino, S. Huet, F. Lefèvre et G. Linarès (2012). Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique. *Proceedings of the national conference "Traitement Automatique des Langues Naturelles" (TALN'12)*, pages 527–534, Grenoble, France.

[Simard et al. 2005] Michel Simard, N. Cancedda, B. Cavestro, M. Dymetman, E. Gaussier, C. Goutte, K. Yamada, P. Langlais et A. Mauser (2005). Translating with non-contiguous phrases. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 755–762, Vancouver, British Columbia, Canada.

[Simard et al. 2007a] Michel Simard, C. Goutte et P. Isabelle (2007a). Statistical Phrase-Based Post-Editing. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies (NAACL-HLT'07)*, pages 507–515, Los Angeles, USA.

[Simard et al. 2007b] Michel Simard, N. Ueffing, P. Isabelle et R. Kuhn (2007b). Rule-Based Translation with Statistical Phrase-Based Post-Editing. *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT'07)*, pages 203–206, Prague, Czech Republic.

[Snover et al. 2006] Matthew Snover, B. Dorr, R. Schwartz, L. Micciulla et J. Makhoul (2006). A study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA'06)*, pages 223–231, Cambridge, USA.

[Snover et al. 2009] Matthew Snover, N. Madnani, B. Dorr et R. Schwartz (2009). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. Machine Translation, 23(2-3) :117–127.

[Snow et al. 2008] Rion Snow, B. O'Connor, D. Jurafsky, et A. Ng (2008). Cheap and Fast – but is it Good ? Evaluating Non-expert Annotations for Natural Language Task. *Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pages 254–263, Honolulu, Hawaii.

[Soricut et al. 2012] Radu Soricut, N. Bach et Z. Wang (2012). The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 145–151, Montréal, Canada. Association for Computational Linguistics.

[Soricut et Echihabi 2010] Radu Soricut et A. Echihabi (2010). TrustRank : Inducing Trust in Automatic Translations via Ranking. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 612–621, Uppsala, Sweden.

[Soricut et Narsale 2012] Radu Soricut et S. Narsale (2012). Combining Quality Prediction and System Selection for Improved Automatic Translation Output. *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT'12)*, pages 163–170, Montréal, Canada. Association for Computational Linguistics.

[Specia 2011] Lucia Specia (2011). Exploiting Objective Annotations for Measuring Translation Post-editing Effort. *proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT'11)*, pages 73–80, Leuven, Belgium.

[Specia et al. 2010] Lucia Specia, N. Cancedda et M. Dymetman (2010). A dataset for Assessing Machine Translation Evaluation Metrics. *7th Conference on International Language Resources and Evaluation (LREC-2010)*, pages 3375–3378, Valletta, Malta.

[Specia et al. 2009a] Lucia Specia, M. Turchi, N. Cancedda, M. Dymetman et N. Cristianini (2009a). Estimating the Sentence-Level Quality of Machine Translation Systems. *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.

[Specia et al. 2009b] Lucia Specia, Z. Wang, M. Turchi, J. Shawe-Taylor et C. Saunders (2009b). Improving the Confidence of Machine Translation Quality Estimates. *Proceedings of the 12th Machine Translation Summit (MT Summit XII)*, pages 26–30, Ottawa, Ontario, Canada.

[Stolcke 2002] Andreas Stolcke (2002). SRILM, an Extensible Language Modeling Toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP- INTERSPEECH'02)*, volume 3, pages 901–904, Denver, Colorado.

[Suzuki 2011] Hirokazu Suzuki (2011). Automatic Post-Editing based on SMT and its selective application by Sentence-level Automatic Quality Evaluation. *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 156–163, Xiamen, China.

[Tillmann et al. 1997a] C. Tillmann, S. Vogel, H. Ney, H. Sawaf et A. Zubiaga (1997a). Accelerated DP based Search for Statistical Translation. *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.

[Tillmann et al. 1997b] C. Tillmann, S. Vogel, H. Ney, H. Sawaf et A. Zubiaga (1997b). Accelerated DP-based Search for Statistical Translation. *Proceedings of European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Grèce.

[Tomas et Casacuberta 2001] Jesus Tomas et F. Casacuberta (2001). Monotone Statistical Translation using Word Groups. *Proceedings of the 8th Machine Translation Summit (MT Summit VIII)*, pages 357–361, Santiago de Compostela, Spain.

[Turian et al. 2003] Joseph Turian, L. Shen et I. D. Melamed (2003). Evaluation of Machine Translation and its Evaluation. *Proceedings of the 9th Machine Translation Summit (MT Summit IX)*, pages 386–393.

[Vidal 1997] E. Vidal (1997). Finite-State Speech-to-Speech Translation. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 111–114, Munich, Germany.

[Vilar et al. 2006] David Vilar, J. Xu, L. F. D'Haro et H. Ney (2006). Error Analysis of Statistical Machine Translation Output. *proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy.

[Wagner et Fischer 1974] Robert A. Wagner et M. J. Fischer (1974). The String-to-String Correction Problem. Communication of the ACM, 21(1) :168–173.

[Ye et al. 2007] Yang Ye, M. Zhou et C.-Y. Lin (2007). Sentence Level Machine Translation Evaluation as a Ranking Problem : one step aside from BLEU. *Proceedings of the 2nd Workshop on Statistical Machine Translation (ACL-StatMT'07)*, pages 240–248, Prague, Czech Republic.

[Zens et Ney 2004] Richard Zens et H. Ney (2004). Improvements in Phrase-Based Statistical Machine Translation. *Proceedings of the Human Language Technology Conference, the North American Chapter of ACL (HLT-NAACL'04)*, pages 257–264, Boston, USA.

[Zhou et al. 2008] Ming Zhou, B. Wang, S. Liu, M. Li, D. Zhang et T. Zhao (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. *proceedings of the 22nd International Conference on Computational Linguistics (CoLing 2008)*, pages 1121–1128, Manchester, United Kingdom.